Localisation des activités économiques et dépendance à l'échelle géographique considérée

Emmanuel Auvray

▶ **To cite this version:**

Emmanuel Auvray. Localisation des activités économiques et dépendance à l'échelle géographique considérée. Economies et finances. Le Mans Université, 2019. Français. NNT : 2019LEMA2004 . tel-02651531

HAL Id: tel-02651531
https://theses.hal.science/tel-02651531

Submitted on 29 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITE BRETAGNE ECONOMIE LOIRE ET GESTION

Le Mans Université

THESE DE DOCTORAT DE

LE MANS UNIVERSITE
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 597
Sciences Economiques et sciences De Gestion
Spécialité : Économie

Par

Emmanuel AUVRAY

Localisation des activités économiques et dépendance à l'échelle géographique considérée

Thèse présentée et soutenue au Mans, le 04/06/2019
Unité de recherche : GAINS - EA 2167
Thèse N° : 2019LEMA2004

Rapporteurs avant soutenance :

Rachel Guillain Professeur des Universités, Université de Bourgogne
Florence Puech Maitre de conférences, Université de Paris-Sud

Composition du Jury :

Examinateurs :	José de Sousa	Professeur des Universités, Université de Paris-Sud
	Nicolas Le Pape	Professeur des Universités, Université de Caen Normandie
	Stéphane Riou	Professeur des Universités, Université Jean Monnet
Dir. de thèse :	Jean-Pascal Gayant	Professeur des Universités, Le Mans Université
Co-dir. de thèse :	Salima Bouayad Agha	Maitre de conférences, Le Mans Université

Le Mans Université n'entend donner aucune approbation ni improbation aux opinions émises dans cette thèse. Ces opinions doivent être considérées comme propres à leur auteur.

Remerciements

Ce travail a été mené grâce à l'aide directe ou indirecte de très nombreuses personnes. Les premières personnes que je tiens à remercier sont mes co-directeurs de thèse. Jean-Pascal Gayant qui a accepté de se lancer dans cette aventure avec moi malgré les difficultés administratives du début. Merci de ne pas avoir abandonné et de l'aide apporté tout au long de ce travail, notamment les conseils pour être plus pédagogique lors de mes présentations. Merci également à Salima Bouayad-Agha qui a rejoint la thèse au cours de la deuxième année. Ses connaissances ont apporté un regard nouveau sur mes travaux. Je lui dois beaucoup pour son aide et sa capacité à démêler mes idées.

Je tiens aussi à remercier Rachel Guillain et Florence Puech pour avoir accepté de rapporter ma thèse, ainsi que José de Sousa, Nicolas Le Pape et Stéphane Riou de faire partie de mon jury. C'est un privilège de profiter de leurs expertises.

Les membres du GAINS ont aussi une part importante dans l'amélioration des travaux, en particulier lors du séminaire annuel pour les doctorants. En premier lieu François Langot qui dirige le laboratoire avec un engagement exemplaire. Son aide personnelle et le soutien aux doctorants sont non négligeables pour mener des thèses à bien. Je le remercie également pour les nombreuses discussions sur la fonction d'enseignant chercheur et la confiance qu'il a témoigné en laissant mener des projets pédagogiques dans ses enseignements.

Je remercie plus globalement les membres du GAINS pour les contacts quotidiens et, pour certains, de la confiance j'ai enseigné leurs matières : Julien Albertini, Vincent Boitier (avec qui les échanges furent trop peu nombreux lors de cette thèse), Arnaud Chéron, Loïc Du Parquet, Xavier Fairise pour les nombreuses discussions, Yves Guillotin, Frédéric Karame, Sébastien Menard, Huyen Nguyen Thi Thanh, Simon Petitrenaud, Martial Phelippe-Guinvarc'h, Solenne Tanguy, Mouloud Tensaout, Ahmed Tritah, Jonathan Vaksmann, Thomas Weitzenblum et Yves Zenou.

Je tiens aussi à remercier Sylvie Blasco et Jéremy Tanguy pour les moments partagés pendant l'organisation et lors des Journées de Microéconomie Appliquée 2017 au Mans.

Je garde également une pensée pour Stéphane Adjemian. Même si ses recherches ne sont pas dans la même thématique que cette thèse, je me souviendrai toujours de lui comme le premier enseignant d'économie que j'ai eu à l'université. Son enseignement inoubliable à des étudiants de

première année a piqué ma curiosité et c'est en partie grâce à lui que l'idée d'une thèse s'est développée.

Les qualités de Thepthida Sopraseuth et Nicolas Le Pape constituent des modèles de clarté dans leurs enseignements, j'essaye modestement de m'en inspirer lorsque je suis devant des étudiants.

Un grand merci à Anne Bucher, Aymen Esselmi, Zao Khai, Sarah Le Duigou, Benedicte Rouland, Pierre-Jean Messe, Kevin Poperl, Benedicte Rouland. Les anciens doctorants et chargés de travaux dirigés à l'université du Maine lorsque j'étais étudiant. Ils m'ont montré que l'âge n'était pas forcément important pour être clair et apporter une aide précieuse aux étudiants.

Ils ont aussi marqué cette thèse par leur passage au laboratoire du Mans et au bureau des doctorants en tant doctorants, ATER ou CTER : Laurent Brembilla, Nicolas Lefebvre, Enareta Kurtbegu, Arthur Poirier, Shaimaa Yassin. Les moments passés resteront dans ma mémoire.

Les discussions très intéressantes, avec des personnes non moins intéressantes, lors des conférences et des colloques ont également apportés aux réflexions et améliorations menées lors de la thèse : Emilie Arnoult, Catherine Baumont, Enora Belz, Jean Bonnet, Clément Carbonnier, Laetitia Challe, Frédéric Chantreuil, Sylvain Chareyron, Noé Ciet, Laurent Denant-Boèmont, Jean-Michel Floch, Ewen Gallic, Jean-Pascal Guironnet, Morgane Laouenan, Marie-Noëlle Lefebvre Etienne Lehmann, Olivier L'Haridon, Yannick L'Horty, Pascale Petit, Vivien Roussez, Sébastien Roux, Florent Sari et Muriel Travers. Je destine également des remerciements particuliers à Laurent Gobillon et Miren Lafourcade pour l'organisation des Regional and Urban Economics Seminar auxquels j'ai assistés.

Je souligne aussi l'importance du personnel administratif et technique de la Faculté de Sciences Économiques de Le Mans Université. Ils sont plus que des rouages dans l'organisation. L'aide et la bienveillance témoignées ne sont pas à négliger pour travailler sereinement. Les étudiants auxquels j'ai eu l'opportunité d'enseigner sont aussi à mentionner. La constante remise en question nécessaire pour leur apporter les meilleures explications et les critiques parfois constructives constituent d'excellents outils de progression.

Il serait difficile de ne pas mentionner Jérôme Ronchetti et Anthony Terriau dans cette section. Leur implication lors de ces années de thèse est allée au-delà d'une simple relation de travail entre doctorants. Ils m'ont donné de nombreux conseils et m'ont permis de gagner un temps précieux sur la partie administrative de la thèse. Merci pour les moments partagés à l'université et en dehors.

Ma famille a évidemment joué un rôle non négligeable par son soutien. Je n'aurais pas pu aboutir à ce parcours sans elle. Isabelle, Sophie et Guillaume, merci à vous ainsi qu'à vos petites familles

pour les moments partagés. Il est très probable que mon parcours universitaire se soit arrêté à la licence sans vous, je vous adresse mes remerciements les plus profonds.

Table des matières

Introduction générale

La loi du 7 août 2015 portant Nouvelle Organisation Territoriale de la République (loi NOTRe) a réorganisé les compétences économiques des territoires. Ce projet de décentralisation avait notamment comme objectif de lutter contre le millefeuille territorial et de clarifier l'identité des interlocuteurs territoriaux. Suite à la réforme, c'est à la région que revient la compétence économique même si des délégations restent possibles, notamment pour des Établissement public de coopération intercommunale (EPCI) ou des communes. Ce changement réduit de manière drastique le rôle du département dans le développement économique territorial.

Cette réorientation des compétences territoriales témoigne de la complexité des schémas de développement économique car chaque territoire peut intervenir selon ses outils propres et des intérêts divergents. De fait, il est très difficile de prendre en compte tous les interlocuteurs territoriaux et leurs décisions pour mesurer l'impact d'une politique particulière. La difficulté est accrue car peu importe l'échelle géographique choisie, des décisions peuvent exister à des échelles plus petites ou plus grandes.

Rappelons tout d'abord que la prise en compte de l'espace n'alla pas de soi dans la littérature économique même si des travaux importants datent de plus d'un siècle. Si le travail de Marshall (1890) est le plus connu, les contributions de l'école germanique sont décisives pour l'intégration de l'espace et de la notion de distance en économie. Ainsi Von Thünen (1826) proposait une première modélisation afin d'expliquer l'allocation optimale des terres à partir du transport des marchandises vers le centre de la ville. Launhardt (1885) modélisa les problématiques de différenciation horizontale et verticale des firmes en parallèle des travaux de Hotelling (1929) [1]. Enfin la recherche de la localisation optimale pour une entreprise fut popularisée par Weber (1909) [2].

Le peu de considération accordé à l'espace peut s'expliquer, à la suite de Isard (1949) par un choix des économistes de porter les efforts de modélisation sur l'aspect temporel plutôt que spatial. Les difficultés liées à la modélisation de l'espace sont aussi évoquées par Thisse (1997). Malgré plusieurs tentatives, notamment d'Alonso (1964) ou Fujita et Thisse (1986), ce sont les travaux de (Krugman, 1991b,a, 1997) qui, de par la modélisation effectuée, ont permis l'intégration de l'espace

1. Voir Pinto (1977) et Ferreira et Thisse (1996) pour plus de précisions sur les travaux de Launhardt (1885).
2. Voir notamment Thisse *et al.* (1984) pour plus de précisions sur la recherche du point optimal de localisation.

comme élément de la recherche académique.

Les déterminants de localisation peuvent aussi bien être recherchés à une échelle inférieure à une commune qu'à une échelle nationale ou quasi-continentale. Les déterminants testés sont en revanche, peu ou prou, toujours les mêmes. Le choix de localisation dépend de variables propres à l'entreprise, de variables non locales exogènes à l'entreprise et de variables locales exogènes à l'entreprise. La première catégorie regroupe les caractéristiques de l'entreprise (secteur d'activité, bien produit, taille de l'entreprise,...). La deuxième catégorie se compose de données géographiquement agrégées (taux de chômage d'un territoire, ...). La troisième catégorie regroupe les données localisées (en un point et non dans une zone) ou la distance entre des observations. De plus, comme montré par Arauzo-Carod (2008) l'échelle géographique choisie peut conduire à des erreurs d'estimations. Aussi bien pour les variables explicatives que pour le phénomène expliqué, les résultats sont différents selon le niveau géographique adopté.

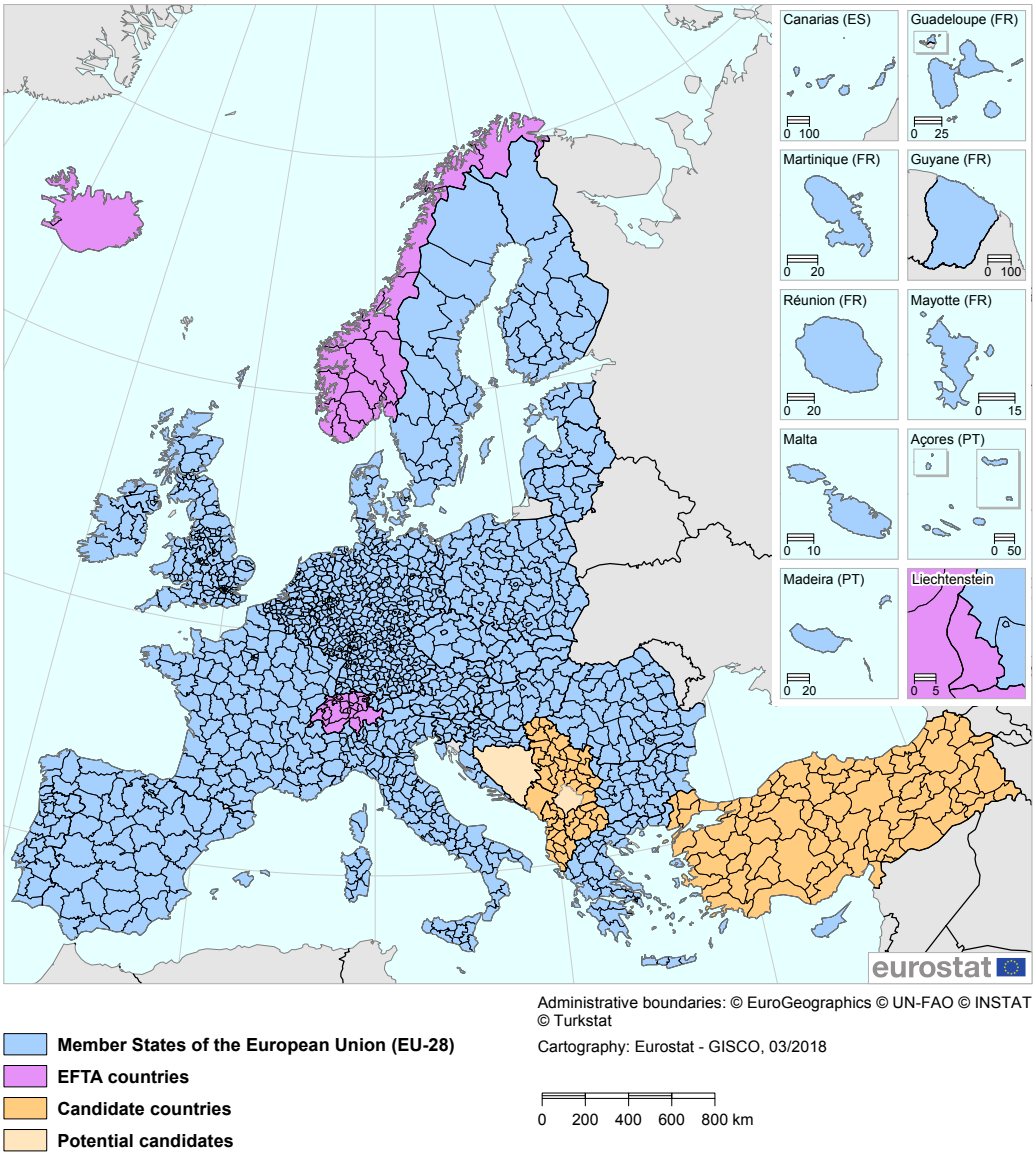
Au cours des différents travaux menés, la question du choix de l'échelle géographique a constitué un élément clé de l'analyse spatiale. Les difficultés à mener des travaux empiriques ont conduit à identifier la problématique dite du MAUP (Modifiable Areal Unit Problem) ; qui désigne le phénomène dépendance des résultats empiriques au choix du découpage géographique retenu (Openshaw et Taylor, 1979).

Les travaux de Briant *et al.* (2010) constituent la meilleure tentative pour mesurer l'effet que pourrait avoir d'autres découpages géographiques sur données françaises. À partir de données géolocalisées, ils appliquent des découpages pour lesquels le nombre de zones est proche des découpages français déjà existants, puis, ils font varier la position de ces zones pour évaluer l'effet zonage du MAUP. Ainsi, les variations entre les estimations peuvent être imputées au MAUP. Cette méthode empirique d'analyse présente l'inconvénient de ne s'intéresser qu'au territoire étudié, l'effet zonage du MAUP n'est calculé que pour la France. L'effet n'est donc pas forcément le même pour d'autre pays.

Les différents zonages rendent difficiles la comparaison du fait de la différence de taille des zones. Les découpages NUTS [3] européens montrent la diversité des découpages entre les différents pays européens. Pourtant, ces découpages sont utilisés à des fin de comparaisons territoriales. La Figure 1 représente le découpage NUTS-3. Visuellement, on constate déjà une forte hétérogénéité entre les pays. Si la superficie des zones est relativement grande pour les pays scandinaves, la superficie est beaucoup plus faible pour l'Allemagne et les pays du Benelux. À titre de comparaison, nous présentons les résultats des découpages NUTS sur la superficie et la population moyenne par zone

3. Nomenclature des Unités Territoriales Statistiques.

FIGURE 1 – Carte du découpage NUTS-3 (2013)



pour l'Allemagne et la France [4] (Tableau 1). Si la superficie est plus importante pour la France, la population est également plus faible. Comme les trois découpages NUTS conduisent à créer plus de zones pour l'Allemagne que pour la France, on pourrait penser que les découpages conduisent à homogénéiser les zones allemandes et françaises, notamment en considérant le découpage NUTS-1 qui permet de comparer des zones aussi peuplées en Allemagne qu'en France. Mais en réalité, plus le découpage est fin, plus les découpages NUTS compliquent la comparaison entre les deux pays. Pour le découpage NUTS-3 par exemple, les zones françaises sont sept fois plus grandes et 3.5 fois plus peuplées. Le découpage NUTS 3 accentue l'hétérogénéité entre les territoires, ce qui rend d'autant plus difficile les comparaisons empiriques. Ce rapide exemple montre l'utilité d'un travail portant sur l'impact du choix découpage géographique pour la mise en œuvre d'études empiriques.

4. Données uniquement pour la France métropolitaine.

Tableau 1 – Détails des découpages NUTS pour l'Allemagne et la France

	Allemagne	France	Écart
Nombre de			
NUTS 1	16	13	-19%
NUTS 2	39	22	-44%
NUTS 3	429	96	-78%
Superficie (en km^2)	348 770	543 965	56%
Superficie moyenne			
NUTS 1	21 798	41 843	92%
NUTS 2	8 943	24 726	176%
NUTS 3	813	5 666	597%
Population	80 767 500	63 697 865	-21%
Population moyenne			
NUTS 1	5 047 969	4 899 836	-3%
NUTS 2	2 070 962	2 895 358	40%
NUTS 3	188 269	663 519	252%

Cette hétérogénéité des découpages pose problème si les zones sont différentes, par leur taille, superficie, etc. Pour les découpages NUTS on peut expliquer cette hétérogénéité par l'utilisation de découpages déjà existants dans chaque pays. On retrouve ainsi la diversité des découpages nationaux qui sont le résultat des histoires respectives de chaque pays et d'organisations territoriales différentes. L'utilisation des découpages NUTS ne permet pas de rendre comparable des territoires situés dans des pays différents.

Afin de prendre en compte l'espace et corriger partiellement les effets du MAUP, une solution possible est de prendre en compte un effet voisinage. Ainsi l'attractivité d'une localisation dépend des caractéristiques associées à la localisation (taux de taxe, réseau de communication,...) mais également des caractéristiques des localisations voisines (ou proches). Cette approche est notamment utilisée pour calculer les indices de Moran (1950) et de Geary (1954). Ces indices permettent l'identification de "pôle chaud" (si les territoires voisins partagent des caractéristiques favorables à la localisation d'activité par exemple), de "pôle froid" (si les territoires voisins partagent des caractéristiques défavorables à la localisation). Il existe également des localisations intermédiaires, s'il n'y a pas de corrélation entre les caractéristiques d'un territoire et les caractéristiques des territoires voisins.

La prise en compte d'un potentiel effet de voisinage est également à la base du modèle autorégressif spatial (SAR) et du modèle avec erreurs spatialement autocorrelées (SEM). Pour estimer ces modèles il est néanmoins nécessaire d'effectuer deux choix importants, le découpage géographique et la matrice de voisinage. Le premier choix peut être relatif à la variable expliquée, dans ce cas il est moins crucial que le second. En effet, la matrice de voisinage choisie est liée à la définition du voisinage des unités spatiales. Cette notion peut dépendre de la contiguïté des zones, de la distance

entre les zones ou du nombre de zones à traverser. Harris *et al.* (2011) et LeSage et Pace (2014) offrent des points de vue différents sur l'utilité et la représentation de la matrice. Si le premier travail considère que le choix de la matrice est capital, le deuxième n'y accorde qu'une importance limitée car les résultats restent robustes malgré plusieurs variantes.

En se focalisant sur les découpages français, nous réduisons l'effet historique pouvant expliquer les différences entre les pays. Le Tableau 2 donne des détails sur différents découpages possibles. Les données françaises nous permettent également de calculer la densité de population par zone. Ainsi on peut vérifier si au sein d'un même pays, les divisions réalisées du territoire permettent d'améliorer les comparaisons entre zones.

On constate globalement que la population et la superficie des zones augmentent au fur et à mesure que le découpage est de moins en moins fin, à l'exception des Arrondissements et des Zones d'emploi. Cette exception s'explique par la nature de ces découpages, un arrondissement est une partie d'un département alors qu'une zone d'emploi peut ne pas faire partie d'un seul département. Les découpages non administratifs (Zone d'emploi et Bassin de vie) permettent d'avoir des frontières qui ne sont pas obligatoirement calquées sur le découpage départemental. Par exemple, une zone d'emploi peut faire partie de plusieurs départements. Ces découpages sont utiles car ils permettent de réellement prendre en compte la zone pertinente pour le chef lieu associé, et compte tenu des autres territoires proches.

Nous pouvons vérifier si les découpages augmentent ou réduisent l'hétérogénéité entre les territoires en s'intéressant plus particulièrement à la densité de population. La densité moyenne par découpage nous indique que la réduction du nombre de zones n'a pas un effet monotone sur la densité moyenne. Ce sont en effet pour des découpages intermédiaires que la densité moyenne est la plus élevée (Arrondissement et Département) alors que la densité moyenne est proche pour les découpages régionaux et le découpage communal. L'écart-type indique néanmoins que les densités sont plus homogènes pour les découpages régionaux que pour des découpages plus fin. On retrouve également que l'hétérogénéité est plus importante quand la densité moyenne est plus élevée.

Les découpages français ne rendent pas comparables des résultats obtenus à différentes échelles. Les découpages non administratifs (Bassin de vie et Zone d'emploi) semblent néanmoins réduire l'hétérogénéité des zones. Cela permet de valider l'utilité de ces zonages spécifiques. Comme les zones sont plus homogènes, la comparaison est plus adaptée.

Afin de mettre en lumière les limites de certaines méthodes du fait du MAUP, on peut évaluer sur données empiriques la variation des résultats suite à une modification du découpage ou de la notion de voisinage. On peut également simuler des localisation de données. Aussi, l'avantage de

Tableau 2 – Détails des découpages français (2016)

	Min	25%	50%	75%	Max	Moyenne	Écart-type
Population							
Commune (35 756)	0	197	441	1 099	2 220 445	1 803	15 127
Bassin de vie (1 642)	1 869	9 492	14 552	23 911.8	10 825 722	39 272.7	283 455.8
Arrondissement (322)	8 004	73 189	129 526.5	246 455.8	2 220 445	200 266.2	224 560.5
Zone d'emploi (306)	11 048	57 465	106 823.5	229 145.5	5 891 175	212 124	413 277.1
Département (96)	76 360	302 540.8	537 246.5	840 473.8	2 603 472	671 726.1	523 143.5
Région (22)	324 212	1 520 171.3	2 136 246	3 326 630.3	12 027 565	2 931 168.3	2 508 981.5
Région 2016 (13)	324 212	3 276 543	5 441 183	5 879 144	12 027 565	4 960 438.7	2 887 927.3
Superficie (en km^2)							
Commune	0	0.03	6.5	10.9	18.7	757.8	15.4
Bassin de vie	9.4	149.6	263.5	429.6	3 960.3	334.3	288.8
Arrondissement	35.4	1 192.4	1 632.7	2 195.8	6 140.1	1 704.7	841.6
Zone d'emploi	119.1	827.3	1 364.6	2 349.6	8 822.2	1 805.6	1 388.5
Département	105.4	5 202.8	5 998.6	6 813.3	10 168	5 717.8	1 946.8
Région	8 324.6	16 486.5	25 836.3	31 812.8	45 587.1	24 950.4	11 381.3
Région 2016	8 726.4	30 137.5	32 395	57 691.5	84 821.7	42 223.8	23 432.8
Densité (en hab/km^2)							
Commune	0	18.5	40	93.3	27126.1	158.6	721.4
Bassin de vie	5.5	35.8	66.6	120.8	2 733.6	108	147
Arrondissement	7.1	42.1	75.9	159.1	21 072.8	442.3	1 740.2
Zone d'emploi	12.6	47.1	77.7	143.1	9 114	181.6	604.1
Département	14.8	49.7	81.6	157.7	21 072.8	563.1	2 460.7
Région	37.2	65.2	91	138.2	997	146.2	201.7
Région 2016	37.2	69.3	110.3	119.1	997	170.4	252

simuler des données géographiques selon un processus totalement, ou partiellement, contrôlé permet de vérifier la pertinence d'une méthode d'analyse au regard du choix du découpage géographique, et ce sans dépendre d'une répartition empirique d'activité. Les méthodes que nous utilisons se distinguent par la création et la simulation de localisations géographiques au lieu de reprendre des données empiriques. Ainsi, nos données simulées ne sont pas dépendantes de spécificités locales ou d'autres phénomènes inobservables. L'identification de biais ou l'évaluation de la pertinence des méthodes ne peut se faire sur données empiriques, sinon, les résultats dépendent d'un territoire spécifique.

Habituellement, l'utilisation de données non empiriques se fait principalement pour introduire le principe du MAUP. Briant *et al.* (2010) schématisent ainsi les effets de zonage et de forme du MAUP avant de l'évaluer sur données françaises. Fratesi (2008) et Guimaraes *et al.* (2011) illustrent les problèmes plus spécifiques liés aux mesures de concentration [5] et à la possibilité d'obtenir des valeurs d'indices égales alors que les logiques de concentration sont différentes. Une autre utilisation possible de ces répartitions géographiques simulées peut se faire pour illustrer des discussions sur les propriétés que les indices de concentration doivent respecter. Thomas-Agnan et Bonneu (2015a) simulent des données pour définir de nouvelles propriétés mais également pour vérifier la validité d'une nouvelle méthode d'évaluation de la concentration spatiale. La présente thèse généralise l'étude des outils appliqués sur des données simulées.

5. Le même terme "indice de concentration" est utilisé dans la littérature pour désigner des concepts différents. Pour couper court à toute ambiguïté, nous précisons dans l'annexe de cette introduction le sens précis de chacun des termes que nous utiliserons.

L'objectif global de cette thèse est d'étudier sur des données géographiquement simulées l'impact du choix géographique effectué. Au lieu d'appliquer des méthodes déjà existantes sur un sujet particulier, nous avons fait le choix d'analyser des méthodes couramment utilisées et de vérifier si elles sont sensibles au découpage géographique choisi. Nous avons axés nos recherches sur l'identification du biais lié au MAUP et la capacité des modèles d'estimation à restituer correctement les déterminants de localisation selon le découpage territorial effectué, puis de proposer une méthode permettant de définir le périmètre géographique pertinent d'un territoire. Ces travaux s'inscrivent notamment dans la lignée de Briant *et al.* (2010) pour mesurer les effets du MAUP, et d'Arauzo-Carod (2008) pour évaluer l'impact du découpage géographique utilisé pour les variables explicatives, sur des résultats économétriques.

Nous nous focaliserons sur les indices de concentration géographique des établissements et les modèles d'estimation utilisés pour identifier les déterminants de localisation. Avec des données simulées, nous pouvons notamment décomposer les effets du MAUP sur les indices concentration dans le premier chapitre, identifier des limites à l'application des indices continus dans le chapitre 2, et évaluer la sensibilité des modèles d'estimation dans le chapitre 3. Cela nous a amené, dans le dernier chapitre, à réfléchir à un outil servant à déterminer le bassin économique local d'un territoire.

⁂

Le chapitre 1 mesure la sensibilité des indices de concentration géographique aux effets taille et forme du MAUP. Le choix est fait de se focaliser sur les indices qui discrétisent l'espace. Même si ces indices sont imparfaits du fait de cette discrétisation ils sont néanmoins simples et couramment utilisés dans la littérature. Afin d'éviter que les résultats soient dépendants du territoire choisi (avec le risque d'omettre des spécificités locales), les localisations d'établissements sont simulées selon des processus contrôlés totalement ou partiellement. L'avantage de cette méthode est que nous pouvons quantifier la variation des indices de concentration, mais également vérifier si les résultats de ces indices restituent les paramètres de localisation. Nous mettons en lumière de manière analytique l'effet d'échelle du MAUP puis nous mettons en évidence l'effet forme en simulant des localisations. Nous effectuons ce travail sur les indices les plus couramment cités, qu'ils soient spécifiques à la concentration géographique des activités (Ellison et Glaeser, 1997; Maurel et Sédillot, 1999) ou issus d'autres domaines (Gini, 1912; Herfindahl, 1950). La méthodologie utilisée rend possible l'évaluation des qualités et défauts d'autres indices. Nos résultats montrent globalement que les indices de Herfindahl et de Gini sont plutôt des indices d'agglomération et non de concentration. De plus, même si l'indice d'Ellison-Glaeser est biaisé par l'agrégation géographique des données,

l'effet forme n'a que peu d'effet sur cet indice. À l'inverse, l'indice de Maurel-Sédillot n'est pas sensible à un changement de l'échelle géographique mais à l'effet forme du MAUP.

Dans le chapitre 2, nous focalisons notre analyse des indices de concentration continus. Ces indices n'utilisent plus l'appartenance à des entités géographiques comme référence mais des indices qui se servent de la distance pour définir la proximité entre les établissements, ces indices sont dits continus car la distance permet d'avoir une approche continue de l'espace. À nouveau nous simulons les localisations des établissements pour contrôler la logique de localisation réelle. Ce travail met en évidence que les choix effectuées pour mesurer la concentration des établissements impactent de manière non négligeable les résultats. Nous nous sommes plus particulièrement intéressés aux localisations contrefactuelles nécessaires pour évaluer si la concentration géographique des établissements des secteurs peut s'expliquer par la localisation des sites qui auraient pues être choisies par les établissements du secteur (et donc de l'agglomération globale des activités économiques). Si oui, il n'y a pas de concentration géographique car l'apparente concentration s'explique en réalité par la géographie des sites possibles. Nous avons également vérifié si la position relative des établissements sur le territoire peut impacter la mesure de concentration, ceci pouvant donc expliquer si des activités frontalières sont plus facilement catégorisées comme concentrés, dispersés, ou ni l'un ni l'autre (c'est-à-dire que la répartition des établissements du secteur correspond à la répartition de l'activité dans son ensemble). Enfin, nous mettons en lumière que le choix du secteur d'intérêt constitue une limite importante des mesures de concentration car l'agrégation sectorielle peut entrainer la réunion de secteurs dont les logiques d'agglomération sont très hétérogènes.

La question du choix du découpage géographique est toujours centrale dans le troisième chapitre mais n'a plus comme sujet principal la concentration des activités. En effet l'objet de ce chapitre est d'étudier la capacité des modèles économétriques à restituer les facteurs déterminants d'un modèle de choix de localisation. L'esprit de ce chapitre est néanmoins dans la continuité des précédents car nous simulons les implantations des établissements d'un secteur au lieu de privilégier des données empiriques. Nos localisations sont simulées à partir d'un modèle de Weber (1909, 1929). Dans ce modèle, la localisation optimale est celle qui minimise le coût de transport total entre tous les fournisseurs d'inputs et le marché final d'output. Ce modèle a l'avantage de dépendre uniquement de la distance et non de l'appartenance à des zones géographiques. De fait, nous pouvons vérifier si en appliquant des découpages géographiques, donc en utilisant des données géographiquement agrégées, on peut retrouver les facteurs déterminants du modèle de localisation. Les différentes spécifications utilisées sont globalement corrélées et sont impactées de la même manière par les modifications successives des paramètres du modèle simulé. Nous trouvons également que les résultats sont sensibles au découpage géographique utilisé, particulièrement quand le nombre de régions est

faible.

Le chapitre 4 constitue une solution au problème de la pertinence des découpages géographiques pour traiter des questions économiques. Nous cherchons quelle est la délimitation du périmètre géographique pertinent pour l'étude d'impact d'une manifestation ou d'une infrastructure. La méthodologie standard de ces études implique que les dépenses des visiteurs locaux ne doivent pas être comptabilisées dans l'évaluation de l'impact de l'événement sans avoir une mesure claire et commune du périmètre géographique pertinent. Hors, les dépenses effectuées par des locaux ne constituent pas une création de richesse pour le territoire car, même en l'absence de la manifestation, les consommateurs locaux auraient dépensé leur argent dans le bassin géographique concerné. L'apport de richesses dû à un événement se limite donc aux dépenses effectuées par des visiteurs qui ne seraient pas venus sans que l'événement soit organisé. Ce travail s'appuie sur les données domicile-travail par commune en France (INSEE [6] 2015), pour identifier les liens existants entre les communes. L'utilisation des flux de travailleurs domicile-travail présente l'avantage de lier les territoires de manière dirigée (c'est-à-dire que l'importance de la ville A pour la ville B n'est pas égale à l'importance de la ville B pour la ville A). De plus, les flux sont associé à l'échelle communale donc les flux sont peu agrégés géographiquement, notre outil nous permet ainsi de ne pas être sensible au MAUP. À l'aide de ces informations, nous développons deux algorithmes d'agglomération des communes à partir d'un lieu de référence. Les bassins ainsi obtenus peuvent être comparés aux différents découpages géographiques existants et servir pour des évaluations de politiques publiques.

6. Institut National de la Statistique et des Études Économiques.

0.A Annexes

Localisation

La localisation désigne la présence d'une firme sur un territoire. La localisation ne désigne l'implantation d'un établissement qu'au sens physique du terme. Aucun traitement statistique n'est effectué pour cette étape.

Agglomération

L'agglomération désigne le regroupement d'activités : il peut s'agir des établissements ou un secteur précis, sur un territoire. Seul un comptage brut des observations est effectué sans mesure de correction ou de pondération.

Concentration

Dans cette thèse, le terme concentration est utilisé pour désigner un forte proximité géographique des établissements d'un même secteur. La concentration est obtenue suite à la correction de l'agglomération des établissements compte tenu du poids des zones. Selon les indices utilisés, on peut dire que les établissements d'un secteur sont concentrés à une certaine échelle géographique ou jusqu'à une certaine distance. Néanmoins, les indices ne permettent pas de connaitre le ou les endroits où les établissements sont concentrés.

Spécialisation

Dans cette thèse, la spécialisation désigne le fait que l'activité d'un territoire soit marquée par l'activité d'un secteur. Il n'y a pas besoin que le secteur soit concentré pour que la zone soit spécialisée, de même, ce n'est pas car un secteur est concentré qu'un ou des territoires sont spécialisés. La spécialisation s'accompagne souvent de la dépendance d'un territoire à une entreprise ou un secteur. Cela peut être avantageux quand le secteur est porteur, mais sur le long terme l'attractivité du territoire sera réduite si le secteur est sur le déclin.

CHAPITRE 1

Les indices de concentration géographique à l'épreuve de l'agrégation des données

Ce chapitre fait l'objet d'un travail co-écrit avec Salima Bouayad Agha (GAINS)

Résumé

Pour caractériser la concentration spatiale des activités économiques il convient de disposer de mesures statistiques fiables afin d'évaluer les disparités existantes et de pouvoir comparer les niveaux de concentration par secteur dans le temps et dans l'espace. L'espace est continu mais sa discrétisation du fait du regroupement spatial d'observations à des échelles géographiques différentes (communes, départements, régions) peut induire une erreur de mesure (Briant *et al.*, 2010). Comme il n'est pas toujours possible de mobiliser la position exacte des entités, ce travail se propose d'étudier, à partir de données simulées, jusqu'à quel point les indices de concentration géographique des activités peuvent être biaisés par l'agrégation géographique. Nous montrons que les valeurs des indices sont sensibles à l'échelle géographique sur la base desquels ils sont calculés et, que certains indices sont plus robustes que d'autres à l'agrégation géographique.

1.1 Introduction

1.2 Introduction

Les avantages liés à la concentration géographique des entreprises ont été mis en évidence il y a plus d'un siècle par Marshall (1890). Pour lui, ces bénéfices sont favorisés par la proximité géographique des entreprises d'un même secteur et permettent *i)* de réduire les coûts de transport entre clients et fournisseurs, *ii)* de pouvoir disposer d'une main d'œuvre spécialisée et stable et *iii)* de bénéficier d'un phénomène de diffusion des connaissances (externalités technologiques). Plus tard, les travaux de Porter (1990, 1998) et l'étude des *learning regions* (Florida et Smith, 1995) se focalisent sur les effets de l'innovation et de la coopération active en R&D sur l'agglomération. Ils soulignent l'importance de la formation de clusters, tant en termes d'innovation technologique qu'organisationnelle. Ainsi, la formation de clusters à forte intensité en R&D participe à la spécialisation économique de certains territoires. L'agglomération des activités sur certaines régions d'un territoire donné, source de disparités territoriales, permet également d'apprécier le niveau de développement économique entre ces zones et le niveau de convergence de ces régions (Barro et Sala-i Martin, 1992).

Les économies d'agglomération désignent les bénéfices directs ou indirects associés à la densité et à la diversité des agents économiques au niveau local. D'après Hoover (1937), il peut s'agir d'économies d'échelle internes (propres à l'entreprise), d'économies sectorielles (qui bénéficient aux entreprises du secteur même si elles ne sont pas dans la même zone géographique) ou d'économies d'urbanisation (qui profitent aux entreprises d'une zone géographique mêmes si elles ne sont pas du même secteur), les activités économiques ne se répartissent pas uniformément sur le territoire, ce qui peut aboutir à une spécialisation de certains territoires [1]. Dans certaines situations il est possible qu'une concentration des établissements soit le résultat d'une répartition non homogène de l'activité sur le territoire (*contagion apparente*) sans qu'il n'y ait d'interactions entre les établissements du secteur. Dans d'autres cas, la concentration est liée à des interactions importantes entre les agents du secteur (*vraie contagion*), qu'il y ait ou non une répartition homogène des activités (Arbia et Espa, 1996).

Que ce soit pour bénéficier des avantages de la proximité, éviter les inconvénients de la dispersion (ou l'inverse selon la logique de localisation d'un secteur), le choix d'implantation est souvent stratégique pour les entreprises. Leur choix de localisation et ce qui le détermine peut être appréhendé à partir de mesures de concentration des activités du secteur selon un découpage géographique

1. La concentration d'un secteur sur un territoire ne s'accompagne pas nécessairement de la spécialisation du territoire (Aiginger et Davies, 2004).

donné. D'autre part, il est intéressant pour les pouvoirs publics locaux de disposer d'informations sur la concentration et la spécialisation du système productif local ou sur l'importance relative des établissements dans l'activité de la zone, de manière à anticiper les conséquences d'un choc sur l'activité économique, ou encore pour évaluer si l'activité sur le territoire considéré repose sur un petit nombre d'établissements , ou bien si elle est répartie dans de nombreux établissements.

Si sur le plan théorique, l'explication des phénomènes d'agglomération est relativement bien décrite (Fujita *et al.*, 1999; Fujita et Thisse, 2002), les travaux empiriques sont moins avancés (Ellison *et al.*, 2010; Gibbons *et al.*, 2014). Pour étudier la localisation des activités et l'intensité de l'agglomération des activités d'un secteur, on utilise des mesures de concentration géographique généralement présentées sous la forme d'indices. S'il existe de nombreux indicateurs de concentration géographique ou de spécialisation (Kubrak, 2013), le choix de la mesure la plus adaptée doit faire l'objet d'une attention particulière pour éviter les erreurs d'interprétation (Marcon et Puech, 2014), d'autant plus que ces indices de concentration permettent dans certains cas d'évaluer la spécialisation des territoires dans le temps (Houdebine, 1999). Les propriétés généralement souhaitées pour un indice de concentration (Ellison et Glaeser, 1997; Combes et Overman, 2004; Duranton et Overman, 2005; Fratesi, 2008; Thomas-Agnan et Bonneu, 2015b) sont synthétisées dans le Tableau 1.1.

Tableau 1.1 – Principales propriétés attendues des indices de concentration

N°	Propriétés	Ellison et Glaeser (1997)	Combes et Overman (2004)	Duranton et Overman (2005)	Fratesi (2008)	Bonneu et Thomas-Agnan (2015)
1	Doit corriger de l'agglomération de l'activité économique	x		x	x	x
2	Doit corriger de la structure sectorielle (la répartition des effectifs au sein des établissements)	x		x	x	x
3	Ne doit pas être sensible à une modification des frontières ou au choix de l'unité géographique		x	x	x	x
4	Ne doit pas être sensible à une modification de la classification sectorielle		x			
5	Hypothèse nulle de répartition aléatoire conditionnellement à la répartition globale des activités	x	x	x	x	x
6	Doit pouvoir être significativement testé	x [2]	x	x	x	x
7	Doit reposer sur des justifications théoriques en cas de rejet (ou non) de l'hypothèse nulle		x			x
8	Doit permettre la comparaison entre les secteurs	x	x	x	x	x
9	Doit permettre la comparaison entre des territoires différents	x	x		x	
10	Doit être facilement utilisable compte tenu des données disponibles	x			x	
11	Doit prendre en compte de l'inhomogénéité spatiale des établissements d'un secteur [3]					x

Nous nous focalisons dans cet article sur les propriétés 3 et 9 relatives à la sensibilité des mesures

au choix du découpage géographique. Les travaux de Briant *et al.* (2010) montrent que ces indices qui reposent sur une discrétisation de l'espace peuvent être biaisés et sont sensibles aux découpages et aux niveaux d'agrégation considérés. Par ailleurs, des travaux récents (Billings et Johnson, 2015; Lafourcade et Mion, 2007; Barlet *et al.*, 2013) mettent en évidence que la taille de l'échantillon (et donc de l'industrie) peut avoir un effet : lorsqu'un secteur comporte peu d'établissements l'indice calculé peut amener à conclure, à tort, que le secteur concerné est concentré.

Alors que la décentralisation est au cœur des politiques publiques d'aménagement du territoire, il est donc important de pouvoir déterminer l'échelle géographique la mieux adaptée à l'étude de la concentration ou tout du moins, de vérifier quels sont les indices les moins sensibles aux variations d'échelle. On notera que les indices proposés par Marcon et Puech (2003) et Duranton et Overman (2005) prennent en compte la distance entre les établissements et non plus leur appartenance à une entité géographique (ce sont les indices de troisième génération). Sur le territoire appréhendé comme un espace continu, ces indices de troisième génération ne prennent pas en compte les découpages géographiques, neutralisant ainsi l'effet d'échelle. En revanche, calculés à partir de données à une échelle désagrégée, ces indices sont plus coûteux en temps de calcul et continuent de présenter un biais négatif lié au nombre d'établissements (Barlet *et al.*, 2013).

Généralement, les valeurs des indices calculés à partir de données agrégées à une échelle géographique spécifique dépendent étroitement des découpages considérés. Sous l'acronyme MAUP [4] on désigne l'influence du découpage spatial sur les résultats de traitements statistiques ou de modélisation. Plus précisément, les formes irrégulières et les limites des maillages administratifs qui ne reflètent pas nécessairement la réalité des distributions spatiales étudiées sont un obstacle à la comparabilité des unités spatiales inégalement subdivisées. Les indices de concentration utilisés dans la littérature ne font pas exception aux mesures dépendantes de l'échelle géographique adoptée. En effet, pour un même secteur, la concentration mesurée à l'échelle communale et départementale peut donner lieu à des conclusions différentes.

Selon Openshaw (1984), le MAUP est une combinaison de deux problèmes distincts mais proches :

— Le problème de l'échelle (*scale effect*) est lié à une variation de l'information engendrée lorsqu'un jeu d'unités spatiales est agrégé afin de former des unités moins nombreuses et plus grandes pour les besoins d'une analyse ou pour des questions de disponibilité des données. Dans ce cas, si le nombre de zones est trop faible celles-ci seront trop homogènes tandis qu'un nombre de zones trop élevé augmente le risque de ne pas disposer, sur celles-ci, d'observations sur les variables d'intérêt. Cependant, augmenter ou diminuer le nombre de régions qui

4. Modifiable Areal Unit Problem (Openshaw et Taylor, 1979, 1981).

découpent le territoire traduit l'effet d'échelle uniquement si les zones font parties d'une logique d'imbrication les unes dans les autres. Si ce n'est pas le cas, cela ne relève pas de l'effet d'échelle du MAUP.

— Le problème de l'agrégation (ou de zonage) (*zone effect* ou *shape effect*) est lié à un changement dans la diversité de l'information engendré par les différents schémas possibles d'agrégation à une même échelle. Cet effet est caractéristique des découpages administratifs (particulièrement électoraux) et vient s'ajouter à l'effet d'échelle.

Afin de souligner l'importance qu'il convient d'accorder à cette question, ce travail se propose d'étudier la sensibilité des indices de concentration les plus couramment utilisés aux deux effets du MAUP. L'originalité de cette contribution tient à la distinction et à l'évaluation de chacun de ces effets. Tout d'abord, à partir des résultats d'une étude empirique, nous mettons les indices à l'épreuve de l'effet d'échelle. Puis, pour mettre en lumière la sensibilité des indices à l'effet de zonage, nous simulons différentes configurations de localisation en faisant varier systématiquement des frontières de manière à disposer de différents découpages d'un même espace. Ensuite, nous étudions la capacité des mesures standards de concentration à restituer ces schémas préalablement définis. Nous montrons que les résultats sont divergents et que certains indices sont plus robustes que d'autres à la manière dont l'espace considéré est divisé. Mises bout à bout, ces deux approches nous permettent d'évaluer la sensibilité des indices de concentration au MAUP. De plus, nous pouvons également évaluer la sensibilité de l'effet de zonage au nombre de régions qui découpent le territoire.

La seconde section présente une revue des indices de concentration les plus souvent mobilisés pour quantifier les niveaux de concentration des activités. La troisième partie analyse le problème d'échelle lié au MAUP sur les indices de concentration. Après la présentation de résultats sur données françaises, nous démontrons analytiquement l'effet de l'agrégation géographique sur les indices de concentration. Dans une quatrième section, afin d'étudier le problème de zonage lié au MAUP, nous présentons les différents schémas de localisation retenus et le protocole général de simulation des données avant de présenter les résultats de l'analyse de sensibilité. Puis dans une dernière partie nous concluons sur l'importance du choix de découpage et l'utilisation des indices de concentration.

1.3 Mesurer la concentration spatiale : une revue des principaux indices

La littérature sur les indices de concentration considère deux générations d'indices. Ceux de première génération (Gini, 1912; Herfindahl, 1950) ne sont pas spécifiques à l'étude de la concentration des activités et sont utilisés en économie géographique alors qu'ils ont initialement été élaborés pour mesurer les inégalités de revenu et la concentration (non géographique) des entreprises sur un marché. Les indices de deuxième génération (Ellison et Glaeser, 1997; Maurel et Sédillot, 1999), prennent en compte l'agglomération de l'ensemble des activités pour mesurer la concentration d'un secteur ainsi que la structure du secteur d'activité (la répartition des emploi au sein des établissements). Ils permettent de vérifier si la concentration des établissements est due à la répartition géographique de l'ensemble des activités sur l'ensemble du territoire. Un secteur est dit concentré si le secteur est sur représenté sur certaines zones du territoire. Ces indices sont dits discrets car ils reposent sur une discrétisation prédéfinie du territoire : l'espace analysé est divisé en plusieurs sous-territoires distincts (comme le découpage régional ou départemental) et le positionnement relatif des zones les unes par rapport aux autres ainsi que les interactions spatiales que cela peut engendre ne sont pas pris en compte. Les indices présentés ici sont ceux les plus couramment utilisés dans la littérature.

1.3.1 Indices discrets de première génération

Indice de Gini

L'indice de Gini est fondé sur une comparaison entre la répartition observée des effectifs salariés sur N régions et une valeur hypothétique parfaitement égalitaire (équirépartition) des effectifs salariés sur ces N régions. Son calcul repose tout simplement sur la courbe de Loren qui permet de mesurer la répartition des effectifs salariés d'une régions dans les établissements du secteur. Pour un secteur k donné, on désigne respectivement par S_k et S^i la part des effectifs du secteur k et de la région i au niveau agrégé. On note également $s_k^i = \frac{\text{Effectif du secteur k dans la zone i}}{\text{Effectif total du secteur k}}$ la part de l'effectif du secteur k dans une région i. On définit r_k^i comme l'écart entre l'effectif constaté (s_k^i) et l'effectif théorique (S^i), soit :

$$r_k^i = s_k^i - S^i$$

Dans le cas où cet écart est nul ($s_k^i = S^i$) pour tout $i \in 1, 2, ..., N-1, N$, la répartition du secteur k correspond à la localisation de l'ensemble des activités au niveau agrégé. Dans le cas contraire ($s_k^i \neq S^i$ pour au moins une valeur de $i \in 1, 2, ..., N-1, N$), la répartition du secteur

k ne correspond pas à la localisation de l'ensemble des activités au niveau agrégé. La courbe de Lorenz associée, notée R_k^i, n'est autre que le cumul des r_k^i triés de manière croissante. L'indice de Gini est calculé comme la somme des écarts entre la répartition homogène et la courbe de Lorenz :

$$G_k = 1 - 2 \times \sum_{i=1}^{N} (R_k^i) \tag{1.1}$$

Si la répartition observée correspond à une équirépartition, on a alors $G_k = 0$ ($r_k^i = 0 \; \forall i$). Plus G_k augmente et plus on s'éloigne d'une équirépartition (avec un maximum de $G_k = 1$ [5]).

Il n'existe pas de valeur de référence qui permettrait de caractériser une localisation d'établissements qui ne serait ni concentrée, ni aléatoire. De plus, des répartitions distinctes dans l'espace peuvent conduire à des valeurs identiques de l'indice, ne permettant pas de se fonder sur la valeur de cet indice pour hiérarchiser les niveaux de concentration d'un secteur. Une variante de cet indice consiste à pondérer la part relative d'un secteur sur un territoire, par le poids que celui-ci occupe sur le territoire total (par exemple, on pondère la part d'un secteur sur un département, par le poids du département sur le territoire national). On parle alors d'indice de Gini relatif (WGini). Cette variante permet de tenir compte dans le calcul de l'importance relative d'une zone donnée dans le calcul de l'indice.

Indice de Herfindahl

Si l'indice de Gini permet de mesurer la concentration par analogie avec la mesure des inégalités, l'indice de Herfindahl permet de mesurer la concentration par analogie avec ce qui est utilisé dans le contexte de la concentration industrielle. Ici, il permet de comparer la répartition des effectifs dans chacun des secteurs à partir d'un découpage en N régions. Si on reprend les mêmes notations que précédemment celui-ci est égal à :

$$H_k = \sum_{i=1}^{N} (s_k^i)^2 \tag{1.2}$$

Pour $H_k = 1$, l'intégralité des effectifs d'un secteur se situe dans une seule région : il y a donc une concentration parfaite du secteur et tous les effectifs (donc les établissements) se localiseront dans cette région. Lorsque $H_k = 1/N$, les effectifs du secteur k sont répartis de manière homogène entre les N régions, ce qui correspond à une dispersion parfaite des effectifs [6]. Dans ce cas, à l'inverse

5. On considère parfois que $G_k = 0.5 - \sum_{i=1}^{N} (R_k^i)$; G_k est alors compris entre 0 et 0.5. Afin de faciliter les comparaisons, nous normaliserons les valeurs pour qu'elles soient bornées entre 0 et 1.

6. Afin de faciliter les comparaisons, nous normaliserons également les valeurs pour qu'elles soient bornées entre

de ce qui se passe dans le cas d'un secteur concentré, la présence d'employés du même secteur à un endroit donné réduit les chances qu'un autre établissement du même secteur se localise à un endroit proche. Lorsque la présence d'employés d'un secteur donné à un endroit donné n'a pas d'impact sur le choix de localisation d'un établissement du même secteur, on admet que la répartition des effectifs sur le territoire est aléatoire. Comme pour l'indice de Gini, on ne peut pas associer une valeur particulière de l'indice à cette configuration spécifique.

1.3.2 Indices discrets de deuxième génération

Les indices de Herfindahl et de Gini ne sont pas des mesures spécifiquement dédiées à la concentration géographique. Ils ne prennent pas en compte l'agglomération globale de l'activité sur les différents territoires. Les indices de seconde génération sont plus pertinents dans la mesure où ils prennent en compte de manière explicite la localisation des établissements.

Indice d'Ellison et Glaeser

L'indice d'Ellison et Glaeser (1997) repose sur une approche probabiliste. Ils définissent un indice de concentration spatiale à partir d'une distribution aléatoire des établissements, selon une probabilité proportionnelle à la taille des zones géographiques. Cet indice repose sur un modèle théorique de choix de localisation dans lequel l'agglomération est expliquée par la présence d'avantages naturels (typiquement un meilleur accès au réseau de communication ou la présence d'une matière première à proximité) et/ou d'externalités du fait de la présence d'autres établissements. Le cadre aléatoire proposé par les deux auteurs permet également de tester la significativité des résultats obtenus.

Dans ce modèle, les M établissements choisissent leur localisation parmi les sites N. Un établissement décide de faire le même choix que l'entreprise précédente ou fait le choix d'une localisation au hasard. À partir de ce modèle, on obtient l'indice d'Ellison et Glaeser qui s'écrit :

$$\gamma_k^{EG} = \frac{\sum_{i=1}^{N}(s_i - x_i)^2 - (1 - \sum_{i=1}^{N} x_i^2)(\sum_{i=1}^{M} z_j^2)}{(1 - \sum_{i=1}^{N} x_i^2)(1 - \sum_{j=1}^{M} z_j^2)} \tag{1.3}$$

où s_i est la part de l'emploi du secteur k dans la zone i, x_i est la part relative de la zone i dans l'emploi total et les z_j mesurent les tailles des établissements j du secteur k.

Les deux valeurs extrêmes de cet indice sont obtenues, comme pour l'indice de Herfindahl, pour l'équirépartition et la concentration totale. Un secteur est dit fortement concentré si la valeur est supérieure à 0.05 et faiblement concentrée si la valeur est inférieure à 0.02. Ces bornes sont

0 et 1.

néanmoins déterminées de manière *ad hoc* à partir des valeurs des secteurs d'activités qui sont reconnus comme concentrés.

On note $H_{Etab_k} = \sum_{j=1}^{N} z_j^2$ qui indique le degré de concentration des emplois du secteur k dans les établissements du secteur. On pose également C_{EG}, la concentration mesurée sans prendre en compte de la structure du secteur (qui est prise en compte par H_{Etab_k}), de la manière suivante :

$$C_{EG} = H_{Etab_k} + \gamma_k^{EG}(1 - H_{Etab_k})$$

Avec :

$$C_{EG} = \frac{\sum_i (s_i - x_i)^2}{1 - \sum_i x_i^2}$$

Si $\gamma_k^{EG} = 0$, alors la répartition des établissements du secteur correspond à la répartition globale des activités. γ est "un excès" de concentration pure (G) par rapport à la concentration de la production (H).

Indice de Maurel et Sédillot

Dans le prolongement de ces travaux, Maurel et Sédillot (1999) proposent un indice modifié à partir d'un autre modèle de choix de localisation. Cet indice s'écrit sous la forme suivante :

$$\gamma_k^{MS} = \frac{\frac{\sum_{i=1}^{N} s_i^2 - \sum_{i=1}^{N} x_i^2}{1 - \sum_{i=1}^{N} x_i^2} - H_{Etab_k}}{1 - H_{Etab_k}} \tag{1.4}$$

Plus généralement, les indices Ellison et Glaeser (EG) et Maurel et Sédillot (MS) peuvent se mettre sous la forme :

$$\gamma_k^m = \frac{C_m - H_{Etab_k}}{1 - H_{Etab_k}}, m \in \{EG, MS\} \tag{1.5}$$

et

$$C_{MS} = \frac{\sum_i s_i^2 - \sum_i x_i^2}{1 - \sum_i x_i^2} \tag{1.6}$$

Les deux mesures prennent en compte l'agglomération des firmes, pour corriger la mesure de concentration d'un secteur (suivant le principe qu'il est attendu de trouver plus d'entreprises correspondant à un certain critère, dans une zone où il y a plus d'entreprises). Devereux *et al.* (2004) soulignent qu'il y a une nuance entre le s_i d'EG et de MS. EG corrige de l'agglomération du secteur industriel dans son ensemble, alors que MS corrige de l'agglomération totale de l'emploi. La différence la plus notable entre les deux indices réside dans le calcul de l'écart entre le poids du secteur dans la zone par rapport au poids de la zone. Alors que, EG prend en compte la somme

des écarts pour chaque zone géographique i, MS considère ces écarts sur l'ensemble du territoire. De ce fait, l'indice MS ne prend pas en compte les spécificités pour chaque zone. Les bornes qui servent à identifier les secteurs concentrés sont les mêmes que pour l'indice EG, tout en admettant que ces valeurs sont choisies de manière arbitraire.

1.4 Sensibilité des indices à l'effet d'échelle

1.4.1 Application sur données françaises

Pour illustrer notre propos nous avons calculé les indices de concentration présentés dans la section précédente afin d'analyser la répartition des activités en France [7]. Les données sont issues du fichier des stocks d'établissements en France au 31 décembre 2014 accessible sur le site de l'INSEE [8]. Nous restreignons notre étude à la France métropolitaine (hors Corse) pour garantir la continuité du territoire et permettre ainsi de garder des logiques de localisation et de concentration similaires. De la même manière, Arbia *et al.* (2008) suppriment la Sicile pour étudier la localisation des activités italiennes pour ne pas fausser les résultats obtenus à partir de la distance euclidienne. Les activités sont localisées à la commune, ce qui nous permet également d'agréger directement les données par arrondissement, département, ancienne région [9], nouvelle région [10]. Nous excluons volontairement les découpages géographiques qui ne recouvrent pas totalement le territoire. Les découpages "Bassin de vie" et "Zone d'emploi" sont également exclus car deux communes d'un même bassin de vie/d'une même zone d'emploi peuvent ne pas appartenir à un même département et ne s'inscrivent pas dans l'imbrication des découpages français. De ce fait, l'effet de l'agrégation pour les découpages "Zone d'emploi" et "Bassin de vie" ne peut se faire qu'en comparaison du découpage communal.

Les résultats de l'agrégation géographique sont présentés dans le Tableau 1.2. Ils indiquent que, pour un même indice, l'effet de l'agrégation géographique des données est identique quel que soit le passage du niveau désagrégé au niveau agrégé immédiatement supérieur ("Commune" à "Arrondissement", "Arrondissement" à "Département", etc). On retrouve des résultats identiques quelle que soit la classification NAF (nomenclature d'activités française) considérée. Comme cela est synthétisé dans le Tableau 1.2, l'agrégation géographique des données se traduit par une baisse systématique de la valeur des indices de Gini et de Gini pondéré : pour un secteur donné, plus le découpage géographique est à une échelle fine, plus la valeur de ces deux indices augmente. Le

7. Il s'agit d'un exercice qui se rapproche de celui réalisé par Briant *et al.* (2010).
8. Institut National de la Statistique et des Études Économiques.
9. Régions existantes entre 1970 et 2015.
10. Régions effectives depuis 2016.

résultat obtenu pour l'indice pondéré est tel que la pondération des zones ne corrige pas le biais de l'indice de Gini. À l'inverse, l'agrégation géographique augmente la concentration mesurée par l'indice de Herfindahl[11] et pour les indices de seconde génération : pour un secteur donné, plus l'échelle du découpage géographique est agrégée et plus le secteur semble concentré.

Les variations dues au découpage géographique étaient attendues car les logiques de localisation ne sont pas les mêmes selon les secteurs considérés la proximité géographique n'est pas identique selon les secteurs. En revanche, l'effet apparemment systématique de l'agrégation des zones géographiques données ainsi qu'un sens de variation différent selon l'indice n'étaient pas attendus.

Pour s'assurer que ce résultat n'est pas exclusivement lié à la nature particulière des données mobilisées, nous étudions dans ce qui suit les propriétés théoriques de l'effet de l'agrégation géographique des données sur chacun des indices de concentrations considérés.

Tableau 1.2 – Concentration moyenne des secteurs par indice et par découpage

	Commune	Arrondissement	Département	Région	Nouvelle région	Bassin de vie	Zone d'emploi
Gini	0.96890	0.69756	0.57698	0.51780	0.42725	0.87655	0.73647
WGini	0.85510	0.69544	0.55111	0.46351	0.36826	0.76159	0.71277
Herfindahl	0.01527	0.03722	0.04260	0.09226	0.08451	0.07701	0.05589
EG	0.00590	0.02056	0.02547	0.04507	0.04892	0.02882	0.02699
MS	0.00581	0.02001	0.02513	0.04505	0.04841	0.02951	0.02642

1.4.2 Approche analytique de l'agrégation géographique

Pour chacune des cinq mesures précédentes nous étudions l'effet de l'agrégation géographique des données (de la fusion de deux juridictions à un échelon donné) et ce que cela peut avoir comme conséquence sur la validité des comparaisons selon les découpages géographiques considérés. Pour cela nous comparons les expressions analytiques des différentes mesures précédentes pour deux découpages géographiques, l'un en N régions et en K secteurs d'activité et l'autre avec $N-1$ régions (donc plus agrégé) suite à la fusion de deux régions.

Indice de Gini

Dans le cas de cet indice, l'agrégation peut potentiellement modifier l'ordre de classement des régions selon leur taille croissante Ainsi, dans le cas de trois régions, l'agrégation des deux plus petites peut être telle que la région nouvellement formée soit, ou non, celle qui a la plus grande taille. Nous avons testé les différentes configurations d'agrégation envisageables en considérant trois, puis quatre régions.

11. La diminution de la valeur moyenne de concentration de l'indice de Herfindahl du découpage "Région" à "Nouvelle Région" est uniquement due à la modification de la valeur de l'indice de Herfindahl en prenant en compte le nombre de zones.

L'indice de Gini du secteur k est :

$$G_k(N,K) = 1 - 2 \times \sum_{i=1}^{N}(R_k^i)$$

ou encore :

$$G_k(N,K) = \frac{2\sum_{i=1}^{N} iE_i^k}{N\sum_{i=1}^{N} E_i^k} - \frac{N+1}{N}$$

avec E_i^k le nombre d'emplois du secteur k dans la région i.

L'objectif est de savoir si $G_k(N,K) > G_k(N-1,K)$. Ce qui revient à vérifier si :

$$\sum_{i=1}^{N}(2(N-1)i+1)E_i^k > 2N\sum_{i'=1}^{N-1} E_{i'}^k$$

avec $E_{i'}^k$ le nombre d'emplois du secteur k dans la région i' issue du découpage géographique agrégé [12].

En procédant aux différentes agrégations géographiques possibles avec deux, trois puis quatre régions, on montre que l'indice de Gini diminue suite à l'agrégation géographique si ces conditions sont vérifiées :

— on agrège l'ensemble des régions.

— le processus d'agrégation implique un nombre important de régions.

— l'agrégation n'implique pas une homogénéisation de l'activité dans les différentes régions.

— la différence de taille, avant la fusion, entre la région la plus grande et la région la plus petite est élevée.

Indice de Herfindahl

L'indice de Herfindahl du secteur k s'écrit :

$$H_k(N,K) = \sum_{i=1}^{N} {s_k^i}^2 = \sum_{i=1}^{N}\left(\frac{E_i^k}{E^k}\right)^2 = \frac{1}{(E^k)^2}\sum_{i}^{N}(E_i^k)^2$$

Pour $N = 3$ régions et $K = 3$ secteurs d'activité, nous avons :

$$H_k(3,3) = \frac{(E_1^k)^2 + (E_2^k)^2 + (E_3^k)^2}{(E^k)^2}$$

Si l'on agrège géographiquement les zones 2 et 3 pour former la nouvelle région 2, l'indice est

12. De manière générale, en posant Z le nombre de régions supprimées suite à l'agrégation géographique, la condition s'écrit : $\sum_{i=1}^{N}(2(N-Z)i+Z)E_i^k > 2N\sum_{i'=1}^{N-Z} E_{i'}^k$

le suivant :

$$H_k(2,3) = \frac{(E_1^k)^2}{(E^k)^2} + \frac{(E_2^k + E_3^k)^2}{(E^k)^2} = H_k(3,3) + \frac{2E_2^k E_3^k}{(E^k)^2}$$

Soit de manière générale, on a :

$$H_k(N', K) = \frac{1}{(E^k)^2} \sum_{i}^{N} (E_i^k)^2 + \frac{1}{(E^k)^2} \sum_{i,j=N'}^{N} E_i^k E_j^k \text{ avec } i \neq j$$

$$H_k(N', K) = H_k(N, K) + \frac{1}{(E^k)^2} \sum_{i,j=N'}^{N} E_i^k E_j^k \text{ avec } i \neq j$$

Comme les effectifs sont supérieurs ou égaux à 0, l'inégalité suivante est toujours vérifiée :

$$H_k(N', K) \geqslant H_k(N, K)$$

L'agrégation géographique ne peut donc pas diminuer la valeur de l'indice de Herfindahl. Plus précisément, cette agrégation des entités géographiques augmente la valeur de l'indice à l'exception des cas pour lesquels les effectifs du secteur dans au moins $N-1$ des N régions agrégées sont égaux à 0. Lorsque le regroupement des juridictions ne modifie pas l'indice de Herfindahl, cela signifie qu'au plus une des régions agrégées comporte le secteur considéré. De fait, plus le secteur est présent dans chacune des régions avant l'agrégation, plus l'indice de Herfindahl augmente du fait du regroupement.

Indice de Maurel et Sédillot

On part de l'expression simplifiée :

$$MS_k(N, K) = \frac{C_{MS_k(N,K)} - H_{Etab_k}}{1 - H_{Etab_k}}$$

avec

$$C_{MS_k(N,K)} = \frac{\sum_{i=1}^{N} s_i^2 - \sum_{i=1}^{N} x_i^2}{1 - \sum_{i=1}^{N} x_i^2} = \frac{H_k(N) - H_T(N)}{1 - H_T(N)}$$

où $H_T(N) = \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} \left(\frac{E_i}{E}\right)^2$

L'agrégation géographique diminue la valeur de l'indice si $MS_k(N', K) < MS_k(N, K)$. Comme H_{Etab_k} n'est pas modifié par l'agrégation géographique des données, on a donc $C_{MS_k(N',K)} < C_{MS_k(N,K)}$, si :

$$\frac{\sum_{i,j=N'}^{N} E_i^k E_j^k}{\sum_{i=1}^{N'-1} E_i^k E_j^k} < \frac{\sum_{i,j=N'}^{N} E_i E_j}{\sum_{i=1}^{N'-1} E_i E_j}$$

Cela revient à conclure que dans le cas où l'agrégation géographique augmente la concentration de l'ensemble des activités proportionnellement plus que la concentration du secteur k, alors la valeur de l'indice diminue. Inversement, si l'agrégation géographique augmente la concentration de l'ensemble de l'activité proportionnellement moins que la concentration du secteur k, alors la valeur de l'indice augmente. On ne peut donc pas conclure de manière systématique à l'impact de l'agrégation géographique sur le sens de variation de l'indice MS.

Indice d'Ellison et Glaeser

Indice d'Ellison et Glaeser

Soit :

$$EG_k(N,K) = \frac{\frac{\sum_{i=1}^{N}(s_i - x_i)^2}{1 - \sum_i^N x_i^2} - H_{Etab_k}}{1 - H_{Etab_k}}$$

D'où :

$$EG_k(N,K) = \frac{\frac{H_k(N) + H_T(N) - 2\sum_{i=1}^{N} s_i x_i}{1 - H_T(N)} - H_{Etab_k}}{1 - H_{Etab_k}}$$

Comme H_{Etab_k} est insensible à l'agrégation géographique des données, EG augmente suite à l'agrégation géographique des données si $EG_k(N',K) > EG_k(N,K)$:

$$\frac{H_k(N') + H_T(N') - 2\sum_{i=1}^{N'} s_i x_i}{1 - H_T(N')} > \frac{H_k(N) + H_T(N) - 2\sum_{i=1}^{N} s_i x_i}{1 - H_T(N)}$$

Or, nous avons montré que l'indice de Herfindahl augmente suite à l'agrégation géographique, donc $H_T(N') > H_T(N)$. Si $EG_k(N',K) > EG_k(N,K)$, il suffit de vérifier l'expression suivante :

$$H_k(N') + H_T(N') - 2\sum_{i=1}^{N'} s_i x_i > H_k(N) + H_T(N) - 2\sum_{i=1}^{N} s_i x_i$$

On suppose que l'agrégation géographique est due à la fusion des régions i et j pour former la région i' ait que $H_k(N') - H_k(N) = 2s_i s_j$ et $H_T(N') - H_T(N) = 2x_i x_j$. On peut résumer $EG_k(N',K) > EG_k(N,K)$ à la vérification de :

$$2s_i s_j + 2x_i x_j - 2x_i s_j - 2x_j s_i > 0$$

$$\Leftrightarrow 2\frac{E_i^k E_j^k}{E^k E^k} + 2\frac{E_i E_j}{EE} - 2\frac{E_i E_j^k}{EE^k} - 2\frac{E_j E_i^k}{E^k E} > 0$$

$$\Leftrightarrow EEE_i^k E_j^k + E^k E^k E_i E_j - EE^k E_i E_j^k - EE^k E_j E_i^k > 0$$

En faisant varier N, cette inégalité est toujours vérifier. Nous admettons ce résultat pour l'agrégation géographique de plus de deux zones.

En conclusion, les résultats analytiques peuvent être synthétisés dans le Tableau 1.3.

Tableau 1.3 – Effets de l'agrégation géographique sur les indices de concentration

	Effet empirique	Effet théorique
Gini	↘	↘ sous conditions
Herfindahl	↗	↗
MS	↗	↗ si la concentration du secteur augmente plus que la concentration de l'ensemble des activités
EG	↗	↗

1.5 Sensibilité des indices à l'effet de zonage

1.5.1 Protocole général

Afin d'étudier l'impact de l'effet de zonage du MAUP sur les indices de concentration, nous simulons une distribution de points représentant les coordonnées des localisations d'établissements de plusieurs secteurs, puis à partir de cette répartition, nous simulons une division du territoire pour pouvoir calculer les valeurs des indices de concentration. Plus précisément, nous prenons en compte 2 000 établissements (localisations) se répartissant en cinq secteurs désignés par A, B, C, D et E. Cette répartition se fait en deux temps.

Dans un premier temps, les 1 000 établissements des secteurs A, B, C et D sont répartis selon deux schémas, répartition contrôlée et semi-contrôlée, suffisants pour mettre en lumière la sensibilité des indices et tels que le nombre d'établissements présents dans chaque secteur soit identique dans les deux cas (Tableau 1.4) [13] :

1. Répartition contrôlée : répartition homogène des établissements et affectation sectorielle arbitraire.
2. Répartitions semi-contrôlées : répartitions homogène ou agglomérée de chaque secteur A, B, C et D.

Dans un second temps, 1 000 autres établissements du secteur E sont répartis de manière homogène sur l'ensemble du territoire et ce quel que soit le schéma de localisation retenu pour les

13. Ces effectifs sont issus de la simulation de la répartition contrôlée.

quatre autres secteurs. Dans le cas de la répartition contrôlée, la prise en compte de ce secteur E permet d'éviter qu'il n'y ait qu'un seul type de secteur dans une zone géographique et d'obtenir ainsi une répartition des établissements plus réaliste. Pour les répartitions semi-contrôlées, les indices calculés pour le secteur E servent de valeurs de référence pour étudier comment les logiques de localisation des autres secteurs peuvent avoir une influence et quelle en est la nature.

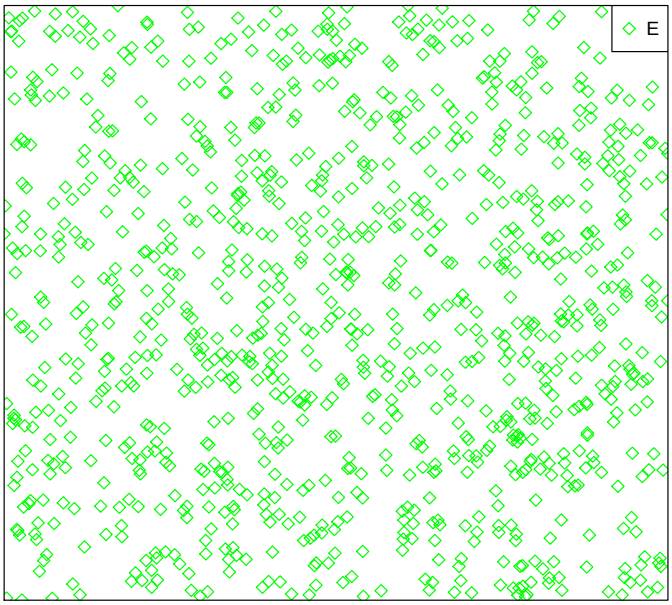


FIGURE 1.1 – Répartition des 1000 établissements du secteur E.

Tableau 1.4 – Nombre d'établissements par secteur

Secteur A	Secteur B	Secteur C	Secteur D	Secteur E
271	231	492	6	1000

Afin d'étudier la sensibilité des indices de concentration à l'effet de zonage du MAUP, nous calculons ceux-ci en faisant systématiquement varier les frontières à l'intérieur d'un territoire donné sur lequel la localisation ou la logique de localisation d'un nombre fixe d'établissements est contrôlée. La définition *a priori* de la localisation des secteurs selon des schémas pré-établis nous permet d'anticiper les valeurs normalement attendues des indices précédemment décrits tandis que la prise en compte de frontières variables nous permet d'étudier l'impact de ces partitions sur la variation des valeurs de ces indices.

La partition du territoire est menée de deux manières distinctes :

1. Une partage en deux zones, soit 17820 découpages différents selon l'inclinaison de la frontière et les aires des zones qu'elle délimite.

2. Un partage variant de deux à cinquante zones en appliquant un diagramme de Voronoï [14].

14. Représente une décomposition particulière d'un espace métrique déterminé par les distances à un ensemble

Pour une même répartition et à nombre de régions donné, nous considérons 100 découpages géographiques. Pour une même répartition d'établissements, cela conduit à découper le territoire 4 900 fois, de telle sorte que les résultats ne dépendent pas d'un découpage unique.

Cette démarche permet d'évaluer la sensibilité des indices de concentration à l'effet de zonage et au nombre de régions qui composent le territoire.

Répartition contrôlée

1 000 établissements (points) sont répartis de manière homogène sur le territoire et sont affectés, selon leur coordonnées géographiques, à chacun des secteurs. Le secteur D se caractérise par une agglomération d'établissements sur une petite aire. Les établissements du secteur B sont localisés de chaque côté de la zone où sont situés les établissements du secteur A, tandis que le secteur C est réparti sur le territoire restant. Ce schéma est représenté par la Figure 1.2 [15].

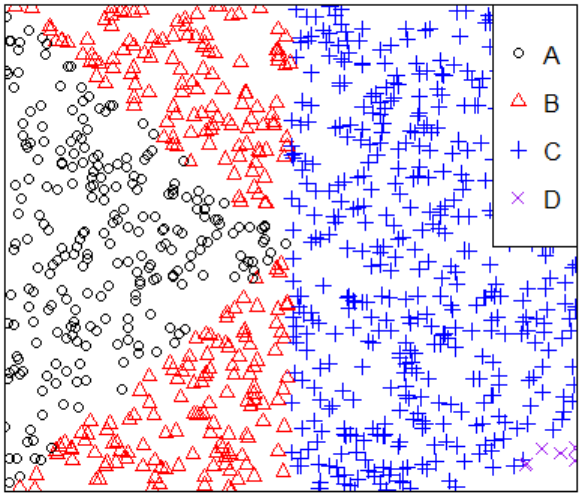


FIGURE 1.2 – Répartition contrôlée des 1000 établissements des secteurs A, B, C et D.

Dans ce contexte, les établissements du secteur E sont répartis de manière homogène sur un territoire où toutes les autres activités sont réparties de manière homogène. Pour ce secteur, par construction, l'indice de concentration varie peu. Si ce n'est pas le cas, et qu'il varie selon les frontières considérées, on dira qu'il est sensible au MAUP.

discret d'objets de l'espace, en général un ensemble discret de points.

15. Répartition homogène et appartenance sectorielle arbitraire.

Répartitions semi-contrôlées

Pour s'assurer que nos résultats ne dépendent pas du schéma pris en compte (contrôlé), nous considérons une autre répartition d'établissements. En conservant les effectifs de chaque secteur (Tableau 1.4) et la localisation des établissements du secteur E, nous modifions la répartition des établissements des secteurs A, B, C et D. Ceux-ci peuvent être localisés de manière homogène ou agglomérée [16]. Cela revient à considérer $2^4 = 16$ combinaisons possibles de logique de localisation pour lesquelles nous simulons une répartition d'établissements [17].

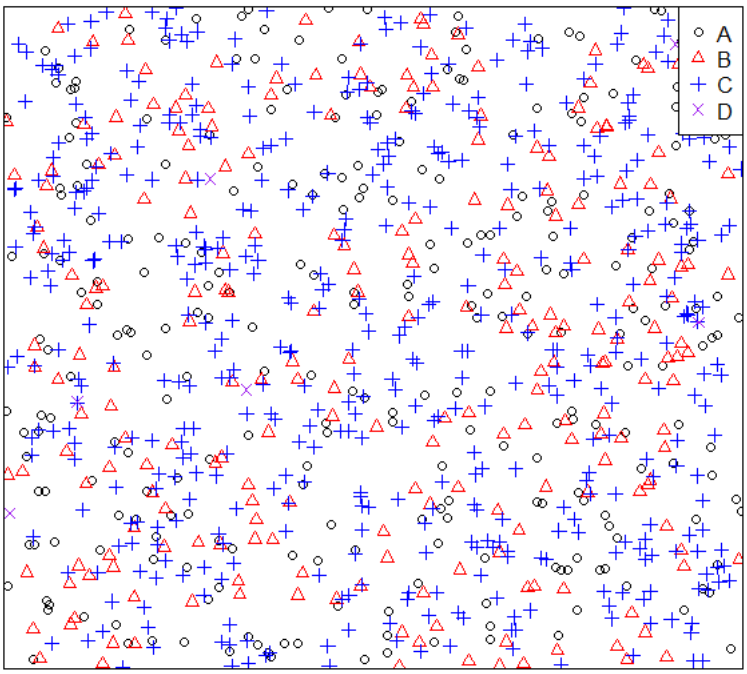


FIGURE 1.3 – Répartition semi-contrôlée des 1 000 établissements des secteurs A, B, C et D si toutes les logiques de localisation sont aléatoires.

Découpages en deux zones

Nous privilégions une analyse graphique à une analyse de corrélation en raison des problèmes liés aux corrélations de Pearson et de Spearman (Hauke et Kossowski, 2011). En effet, des relations non linéaires entre deux variables sont difficilement détectées par un simple de calcul de corrélation. Nous indiquons également la moyenne et l'écart-type de chaque indice, par secteur et par répartition.

16. Pour simuler une répartition agglomérée, nous prenons un processus de (Matérn, 1960) conduisant à une localisation des établissements en un ou plusieurs pôles. Plus précisément, ce processus permet de générer des points (points "fils") dont la localisation dépend de la localisation de points déjà établis (points "pères"). le territoire est représenté par un carré de côté égal à 1. Les paramètres utilisés sont déterminés *ex ante*, la densité du processus de Poisson est de 1, le rayon est de 0.25 et le nombre d'établissements moyen correspond au nombre d'établissements du secteur, de sorte qu'en moyenne il n'existe qu'un seul agrégat de points.

17. Détails en annexe.

Le territoire considéré est un carré de côté égal à 1 de sorte que les coordonnées des établissements varient entre 0 et 1. Cet espace est divisé en deux zones par une frontière linéaire, ce qui permet en modifiant systématiquement son inclinaison de faire varier les aires des zones géographiques prises en compte. Plus précisément, on fait varier l'inclinaison de la frontière de 1° en partant de 0° et en allant jusqu'à 179°. On obtient un ensemble de surfaces variant de 0.01 à 0.99 avec un pas de variation de 0.01. Cela revient à considérer $180 \times 99 = 17820$ frontières différentes délimitant à chaque fois deux surfaces complémentaires sur l'ensemble de l'espace considéré. La construction de cet ensemble de couples de surfaces nous permet de vérifier, à aire égale, la conséquence du choix de la frontière sur les variations des valeurs des indices de concentration, pour chacune des répartitions décrites précédemment. L'effet de la taille et de la forme des zones géographiques est synthétisé par un ensemble de figures en trois dimensions retraçant la forme de la distribution des valeurs des indices selon le degré d'inclinaison de la frontière et la surface des zones. L'effet "forme" du MAUP sera mis en évidence si la valeur de l'indice est sensible au degré d'inclinaison de la frontière pour des tailles de zones données.

En guise d'exemple de lecture de ces graphiques, la Figure 1.4 présente différentes frontières associées à trois surfaces (0.01, 0.5 et 0.99) et trois inclinaisons (0, 90 et 179 degrés). Ces différents points sont représentés dans la Figure 4 où l'axe "Aire" correspond à la surface entre la courbe et la droite d'équation x=1 (soit la frontière Est du cadre) et l'axe "Degré" correspond à l'inclinaison de la frontière. Le Tableau 1.5 synthétise ce à quoi correspondent les neufs points représentés dans la Figure 1.5 [18].

Tableau 1.5 – Configurations type de frontière et d'inclinaison.

	Degré = 0	Degré = 90	Degré = 179
Aire = 0.01	1 ———	2 ———	3 ———
Aire = 0.5	4 – – –	5 – – –	6 – – –
Aire = 0.99	7 · · ··	8 · · ··	9 · · ··

Découpages de Voronoï

Afin de vérifier si les résultats précédents ne sont pas spécifiquement liés à un découpage en deux zones, nous menons le même travail pour des découpages géographiques allant jusqu'à cinquante zones en appliquant un diagramme de Voronoï (Figure 1.6) et en calculant les indices de concentration pour chaque secteur. Pour éviter que nos résultats ne soient sensibles aux diagrammes de

18. La valeur de l'indice de concentration au point 1 de la Figure 1.5 correspond à la frontière pour laquelle l'aire sous la courbe est de 0,01 et le degré d'inclinaison est de 0 ° (Tableau 1.5). La frontière correspondante est la droite ——— représentée dans la Figure 1.4.

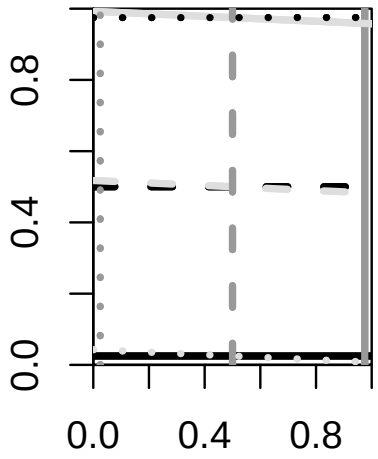


FIGURE 1.4 – Exemples de frontières

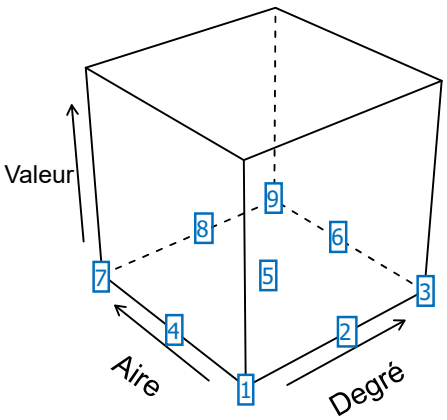


FIGURE 1.5 – Lecture

Voronoï obtenus, nous simulons 100 diagrammes différents avec le même nombre de régions. Cette méthode nous conduit à découper le territoire 4900 fois pour une même répartition d'établissements. Les variations sont directement imputables à l'effet forme du MAUP.

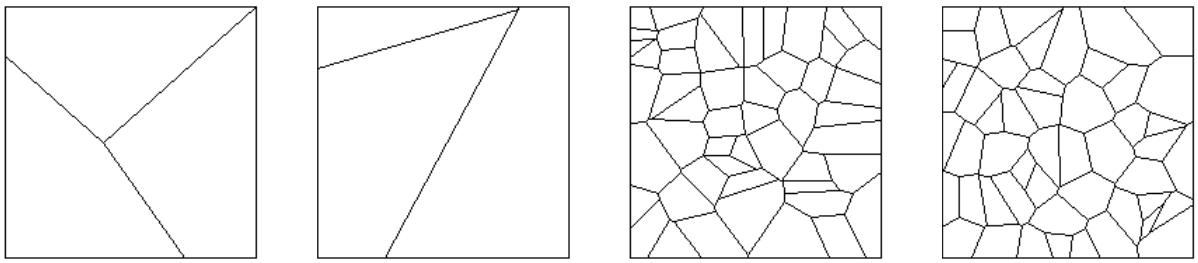


FIGURE 1.6 – Exemples de diagrammes de Voronoï en trois et cinquante zones.

1.5.2 Résultats

Répartition contrôlée

Découpage en deux zones

Le Tableau 1.6 présente la moyenne et l'écart-type de la distribution de chacun des indices par secteur. Comme attendu, le secteur D (par construction) est le plus concentré quel que soit l'indice pris en compte. On constate également que la dispersion de la distribution des indices de seconde génération est la plus importante pour ce secteur. Ces indices semblent donc plus sensibles mais ce résultat est lié au faible nombre d'établissements qui compose ce secteur.

Par ailleurs, pour l'ensemble des indices, le secteur B (après le secteur E) est toujours celui pour lequel la valeur de la concentration est la plus faible, ce qui est conforme à ce qui est attendu, les établissements de ce secteur étant les seuls à être répartis en deux pôles.

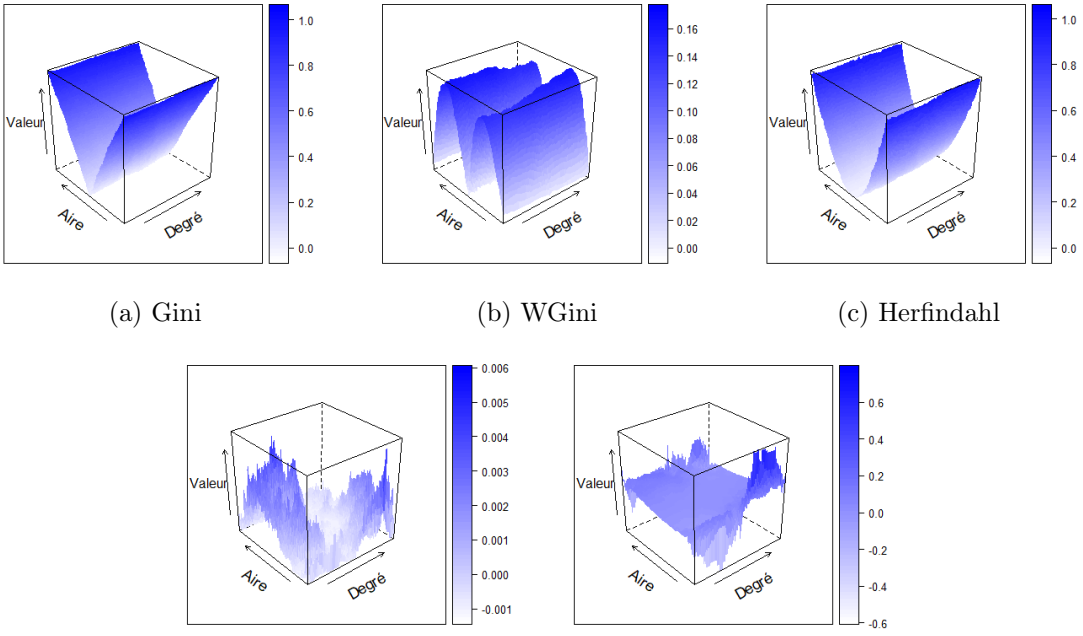
Les établissements du secteur E étant répartis de manière homogène, on s'attend à ce que pour celui-ci les moyennes et les écarts-types de la distribution des indices soient faibles. Cela n'est vérifié que pour l'indice EG (moyenne proche de 0.000 ; écart-type de 0.001). Les indices de seconde génération semblent donc être ceux qui sont les plus conformes aux résultats attendus au vu du schéma de répartition des établissements considéré. De plus, l'indice EG semble plus robuste au MAUP car les écarts-types sont plus faibles qu'ils ne le sont pour l'indice MS.

La lecture des graphiques en trois dimensions suggère que pour le secteur E (Figure 1.7) les indices de première génération sont plus sensibles aux surfaces considérées qu'aux formes des zones définies. Quel que soit le degré d'inclinaison de la frontière, la valeur de l'indice de concentration est déterminée à partir de la surface géographique la plus élevée. De plus, pour les indices de Gini et de Herfindahl, les valeurs sont élevées dès lors qu'il y a plus d'établissements du secteur dans une zone

Tableau 1.6 – Répartition contrôlée : Moyenne et écart-type par indice et par secteur.

	Secteur A	Secteur B	Secteur C	Secteur D	Secteur E
Gini	0.727	0.547	0.620	0.955	0.507
	(0.320)	(0.361)	(0.326)	(0.169)	(0.288)
WGini	0.200	0.115	0.148	0.385	0.099
	(0.156)	(0.010)	(0.113)	(0.270	(0.047)
Herfindahl	0.631	0.430	0.491	0.940	0.340
	(0.387)	(0.400)	(0.379)	(0.213)	(0.299)
EG	0.402	0.125	0.197	2.491	0.000
	(0.422)	(0.167)	(0.204)	(8.00)	(0.001)
MS	0.345	0.138	0.190	0.746	0.025
	(0.779)	(0.624)	(0.647)	(1.861)	(0.078)

géographique. Comme les indices de seconde génération prennent en compte le poids de chaque région pour calculer la concentration, ces indices varient peu selon les différents découpages : ils sont donc les moins sensibles au MAUP.



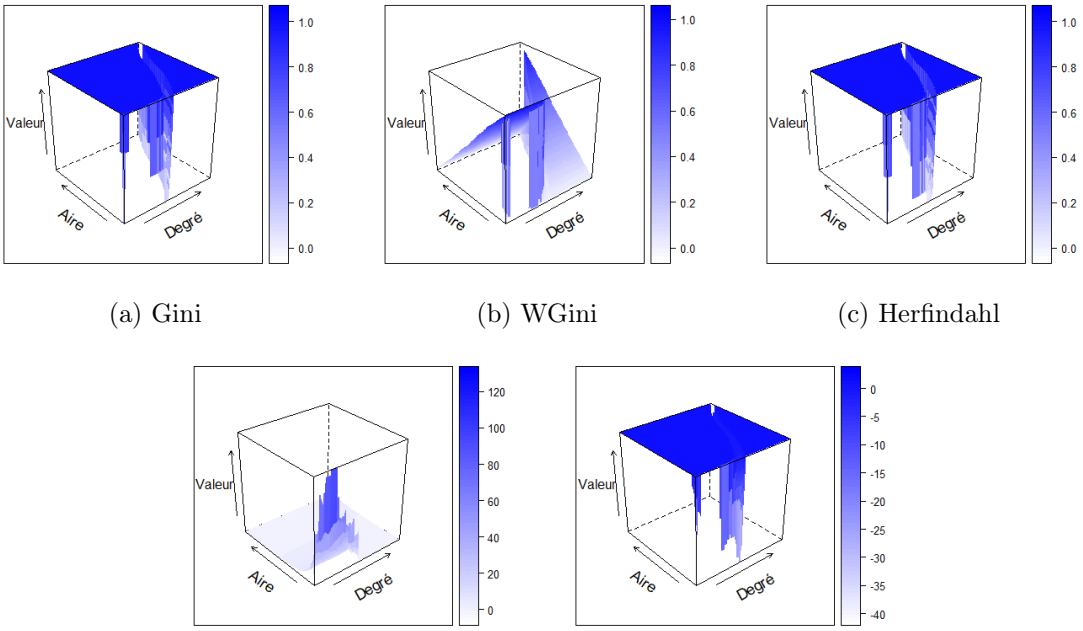
(a) Gini (b) WGini (c) Herfindahl

(d) Ellison-Glaeser (e) Maurel-Sédillot

FIGURE 1.7 – Indices de concentration : Secteur E. Répartition homogène et affectation sectorielle arbitraire.

Pour le secteur D (Figure 1.8), on constate une rupture marquée pour les indices de Gini, Herfindahl et MS. Celle-ci, s'observe à partir d'un seuil pour lequel le dessin de la frontière est tel qu'il y a des établissements du secteur dans les deux zones délimitées par cette frontière. Pour tous les autres découpages pour lesquels les établissements du secteur D ne se localisent que dans une seule zone géographique, la concentration est maximale. L'indice EG prend en compte, l'écart entre la répartition du secteur et la répartition totale, et de ce fait, même si les établissements du secteur D ne se localisent que dans une seule région, la valeur de l'indice de concentration est faible

lorsque la part relative des établissements du secteur D est faible. Le fait que les écarts-types de l'indice MS soient relativement plus élevés que ceux de l'indice EG (à l'exception du secteur D) est essentiellement lié aux valeurs associées aux découpages géographiques les plus spécifiques pour lesquels l'aire de la zone la plus petite est inférieure à 5% du territoire [19].



(a) Gini (b) WGini (c) Herfindahl

(d) Ellison-Glaeser (e) Maurel-Sédillot

FIGURE 1.8 – Indices de concentration : Secteur D. Répartition homogène et affectation sectorielle arbitraire.

Variation des découpages géographiques en N zones

Un test non paramétrique de Kolmogorov-Smirnov permet de vérifier si les distributions des valeurs des indices de concentration selon un découpage en N régions et N+1 régions sont significativement proches. Cela représente 48 tests à indice et secteur donnés car N varie de 2 à 50. Le Tableau 1.7présente le pourcentage des cas où l'hypothèse de similarité des deux distributions est rejetée au seuil de 5% [20].

En moyenne, nous constatons que dans 14% des cas il y a une différence significative entre les distributions des valeurs d'un indice de concentration à N et N+1 régions. L'indice MS est celui qui est le moins influencé par le passage de N à N+1 régions (7% de dissimilarité). La représentation des distributions des valeurs des indices selon le nombre de régions prises en compte (Figures 1.9 et 1.10) permet de vérifier si les variations de chacun des indices sont proches pour un même

19. Les conclusions obtenues à partir des figures A, B et C n'apportent pas d'éléments nouveaux à l'analyse.
20. Pour les tests au seuil de 10% et 1%, les tableaux sont en annexes. Même si les valeurs différent quantitativement, il n'y a pas de différence du fait d'un seuil plus élevé ou plus faible.

Tableau 1.7 – Pourcentage de tests refusés de Kolmogorov-Smirnov à 5%

	Secteur A	Secteur B	Secteur C	Secteur D	Secteur E	Moyenne
Gini	4	13	4	44	2	13
WGini	19	15	17	8	6	13
Herfindahl	13	13	8	58	10	20
EG	19	13	40	8	13	18
MS	10	6	8	4	4	7
Moyenne	13	12	15	25	7	14

secteur. Quel que soit le secteur, c'est pour l'indice MS que la proximité des deux distributions (N et N+1 régions) est la plus forte. On constate, à partir de ce que l'on obtient pour le secteur D, la forte sensibilité de l'indice de Gini et de Herfindahl lorsque le secteur pris en compte comporte un faible nombre d'établissements agglomérés en un coin du territoire (respectivement 44% et 58% de dissimilarités). Pour le secteur C qui présente des localisations d'établissements moins spécifiques, on observe pourtant des différences notables entre les indices, la dissimilarité allant de 7% (Gini) à 40% (EG) contre 15% en moyenne pour ce secteur. Enfin, si l'on admet que le secteur E peut servir de référence, puisque ces établissements sont répartis de manière homogène sur un territoire où l'ensemble des activités est réparti de manière homogène, on peut conclure que, selon ce critère, les indices de Gini, Gini pondéré et MS sont les plus adéquats. L'indice MS semble donc le plus approprié pour comparer des résultats issus de différents découpages géographiques car il est le moins sensible à la logique de localisation du secteur, et montre le moins de dissimilarité.

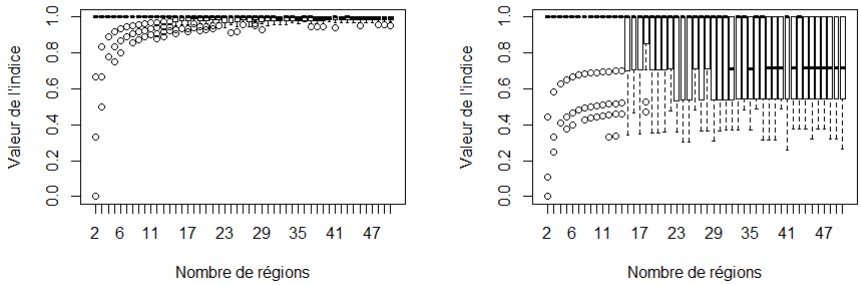


FIGURE 1.9 – Évolution des valeurs de l'indice de Gini et de Herfindahl pour le secteur D

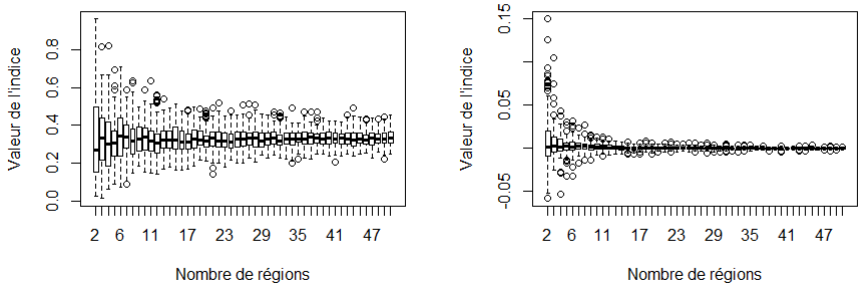


FIGURE 1.10 – Évolution des valeurs de l'indice de Gini et de Maurel-Sédillot pour le secteur E

Variation de la logique de localisation des établissements

Les résultats précédents peuvent être sensibles à la répartition contrôlée des établissements. Pour tester leur robustesse, nous considérons plusieurs autres répartitions d'établissements (dites semi-contrôlées) pour lesquelles seule la logique de localisation de chaque secteur est contrôlée.

On distingue tout d'abord les valeurs des indices uniquement selon la logique de localisation du secteur (Tableau 1.8). On obtient 142560 valeurs d'un même indice pour chaque logique de localisation (huit répartitions aléatoires et huit répartitions agglomérées pour un même secteur).

Cela nous permet de confirmer que l'agglomération augmente la valeur des indices, mais également l'écart-type des indices. En cas de logique de localisation aléatoire, une augmentation du nombre d'établissements d'un secteur fait diminuer la valeur des indices de première génération, fait augmenter la valeur de l'indice EG, mais n'a pas d'effet concluant sur l'indice de MS. Pour une logique de localisation agglomérée, l'augmentation du nombre d'établissements n'a pas d'effet concluant sur les indices à l'exception d'EG pour lequel on observe une légère baisse. En outre, qu'importe l'indice et la logique de localisation, plus il y a d'établissements, plus l'écart-type augmente.

Ce premier exercice mélange néanmoins plusieurs combinaisons de logique de localisation, donc toute chose n'est pas égal par ailleurs. Le Tableau 1.9 indique les valeurs prises quand tous les secteurs sont répartis de manière homogène. Ces valeurs sont proches sauf pour le secteur D pour lequel les écarts-types sont plus élevés. Par ailleurs, les valeurs de concentration sont en moyenne plus élevées pour les indices de première génération et plus faibles pour les indices de seconde génération.

Tableau 1.8 – La valeur des indices de concentration selon la logique de localisation

	Secteur A		Secteur B		Secteur C		Secteur D		Secteur E
	Aléa.	Agglo.	Aléa.	Agglo.	Aléa.	Agglo.	Aléa.	Agglo.	Aléa.
N	142560	142560	142560	142560	142560	142560	142560	142560	285120
Gini	0.498	0.752	0.495	0.673	0.495	0.720	0.533	0.735	0.499
	0.289	0.305	0.288	0.352	0.286	0.332	0.351	0.338	0.289
WGini	0.096	0.170	0.095	0.158	0.095	0.154	0.102	0.180	0.094
	0.048	0.124	0.050	0.139	0.046	0.115	0.078	0.148	0.050
Herfindahl	0.332	0.659	0.328	0.577	0.326	0.629	0.408	0.654	0.333
	0.298	0.376	0.296	0.407	0.292	0.395	0.388	0.415	0.297
EG	0.006	0.243	0.015	0.285	0.003	0.195	-0.056	0.217	0.016
	0.017	0.281	0.027	0.375	0.008	0.219	0.250	0.615	0.028
MS	-0.035	0.510	-0.061	0.328	-0.044	0.482	-0.065	0.393	-0.070
	0.198	0.594	0.247	0.731	0.150	0.571	1.164	0.944	0.222

Le Tableau 1.10 montre l'impact de la concentration d'un secteur lorsque tous les autres secteurs se répartissent de manière aléatoire. Cela ne modifie pas les conclusions précédentes : l'agglomération des établissements d'un secteur augmente la valeur de l'indice ainsi que son écart-type ; moins

Tableau 1.9 – La valeur des indices de concentration si les localisations sont toutes aléatoires

	Secteur A	Secteur B	Secteur C	Secteur D	Secteur E
	Aléa.	Aléa.	Aléa.	Aléa.	Aléa.
N	17820	17820	17820	17820	17820
Gini	0.483	0.493	0.494	0.501	0.499
	0.287	0.293	0.286	0.317	0.289
WGini	0.094	0.096	0.097	0.108	0.098
	0.045	0.046	0.046	0.077	0.046
Herfindahl	0.316	0.329	0.326	0.351	0.333
	0.291	0.302	0.294	0.349	0.297
EG	-0.002	-0.001	-0.001	-0.043	-0.001
	0.003	0.003	0.002	0.237	<0.001
MS	-0.030	0.010	-0.008	-0.248	0.009
	0.132	0.143	0.122	1.260	0.045

il y a d'établissements, plus cette augmentation est importante. On peut donc tester si les distributions des valeurs d'un indice pour un secteur lorsque tous les autres se répartissent de manière aléatoire sont significativement différentes lorsque les établissements de cesecteur sont eux-mêmes répartis de manière aléatoire ou agglomérée. Les résultats du test de Kolmogorov-Smirnov montrent que pour tous les indices les distributions sont significativement différentes au seuil de 1%.

Tableau 1.10 – Impact de l'agglomération des établissements sur les indices de concentration du secteur

	Secteur A		Secteur B		Secteur C		Secteur D		Secteur E
	Aléa.	Agglo.	Aléa.	Agglo.	Aléa.	Agglo.	Aléa.	Agglo.	Aléa.
N	17820	17820	17820	17820	17820	17820	17820	17820	17820
Gini	0.483	0.719***	0.493	0.775***	0.494	0.513***	0.501	0.790***	0.499
	0.287	0.298	0.293	0.310	0.286	0.405	0.317	0.337	0.289
WGini	0.094	0.191***	0.096	0.176***	0.097	0.110***	0.108	0.210***	0.098
	0.045	0.137	0.046	0.121	0.046	0.113	0.077	0.161	0.046
Herfindahl	0.316	0.606***	0.329	0.697***	0.326	0.427***	0.351	0.737***	0.333
	0.291	0.367	0.302	0.380	0.294	0.422	0.349	0.389	0.297
EG	-0.002	0.317***	-0.001	0.231***	-0.001	0.184***	-0.043	0.320***	-0.001
	0.003	0.325	0.003	0.209	0.002	0.174	0.237	0.523	<0.001
MS	-0.030	0.368***	0.010	0.608***	-0.008	0.180***	-0.248	0.555***	0.009
	0.132	0.668	0.143	0.494	0.122	0.661	1.260	0.722	0.045

Comme évoqué précédemment, l'agglomération des établissements d'un secteur donné modifie la répartition globale de l'activité et par conséquent doit modifier les valeurs des indices de concentration de ce secteur mais également celles des autres secteurs. Pour mesurer cet effet, nous considérons le secteur E, réparti de manière homogène quelle que soit la logique de localisation des autres secteurs. Le Tableau 1.11 nous permet de comparer l'effet de l'agglomération d'un seul des quatre secteurs sur les mesures de concentration du secteur E. Comme attendu, il n'y a aucun impact sur les valeurs de l'indice de Gini et de Herfindahl puisque ces indices ne tiennent pas compte de la localisation des autres secteurs. Pour les indices de Gini pondéré et EG, l'agglomération des établissements d'un secteur augmente la concentration mesurée du secteur E ; cette augmentation

est d'autant plus importante que le nombre d'établissements du secteur qui s'agglomèrent est élevé. Il n'y a en revanche pas d'effet systématique sur les valeurs de l'indice MS.

Tableau 1.11 – La valeur des indices de concentration du secteur E selon la logique de localisation des autres secteurs.

	Si A,B,C,D aléa.	Si A agglo.	Si B agglo.	Si C agglo.	Si D agglo.
Gini	0.499	0.499	0.499	0.499	0.499
	0.289	0.289	0.289	0.289	0.289
WGini	0.098	0.100***	0.094***	0.105***	0.097***
	0.046	0.050	0.046	0.055	0.045
Herfindahl	0.333	0.333	0.333	0.333	0.333
	0.297	0.297	0.297	0.297	0.297
EG	-0.001	0.008***	0.005***	0.018***	-0.001***
	<0.001	0.010	0.006	0.018	0.001
MS	0.009	0.008***	-0.042***	0.019***	0.015***
	0.045	0.138	0.085	0.217	0.060

1.6 Conclusion

Le travail précédent illustre l'importance du choix de l'indice adéquat pour rendre compte de la concentration des activités dans l'espace. Nous avons mis en évidence, aussi bien empiriquement qu'analytiquement que l'agrégation géographique entraîne systématiquement une hausse (indice de Herfindahl) ou une baisse (indice de Gini) de la concentration mesurée des indices de première génération. Les indices de seconde génération plus spécifiquement destinés à mesurer la concentration géographique des établissements sont plus adaptés car ils prennent en compte l'agglomération globale de l'activité sur les différents territoires. Nous avons également constaté la pertinence des indices de seconde génération pour prendre en compte les interactions des logiques de localisation de différents secteurs. En faisant varier de manière systématique des frontières partitionnant l'espace sur une même répartition d'établissements, puis en faisant varier le nombre de régions, on constate que l'indice MS est l'indice discret le plus robuste pour restituer les valeurs attendues et celui qui semble être le moins sensible au MAUP malgré une variation plus importante que l'indice EG lorsque les découpages géographiques sont atypiques.

Du fait de cette sensibilité au MAUP, le choix d'un indice selon le découpage géographique considéré n'est pas anodin lorsqu'il s'agit de comparer des secteurs ou lorsqu'il s'agit de les prendre en compte dans des modèles de choix de localisation. De plus, les indices discrets ne permettent pas de savoir si les secteurs concentrés le sont en un ou plusieurs pôles. Il serait intéressant de prolonger cette étude en utilisant des indices de troisième génération.

1.A Annexes

Tableau 1.A.1 – Logiques de localisation des seize répartitions semi-contrôlées

N°	Secteur A	Secteur B	Secteur C	Secteur D	Secteur E
1	Aléatoire	Aléatoire	Aléatoire	Aléatoire	Aléatoire
2	Agglomérée	Aléatoire	Aléatoire	Aléatoire	Aléatoire
3	Aléatoire	Agglomérée	Aléatoire	Aléatoire	Aléatoire
4	Agglomérée	Agglomérée	Aléatoire	Aléatoire	Aléatoire
5	Aléatoire	Aléatoire	Agglomérée	Aléatoire	Aléatoire
6	Agglomérée	Aléatoire	Agglomérée	Aléatoire	Aléatoire
7	Aléatoire	Agglomérée	Agglomérée	Aléatoire	Aléatoire
8	Agglomérée	Agglomérée	Agglomérée	Aléatoire	Aléatoire
9	Aléatoire	Aléatoire	Aléatoire	Agglomérée	Aléatoire
10	Agglomérée	Aléatoire	Aléatoire	Agglomérée	Aléatoire
11	Aléatoire	Agglomérée	Aléatoire	Agglomérée	Aléatoire
12	Agglomérée	Agglomérée	Aléatoire	Agglomérée	Aléatoire
13	Aléatoire	Aléatoire	Agglomérée	Agglomérée	Aléatoire
14	Agglomérée	Aléatoire	Agglomérée	Agglomérée	Aléatoire
15	Aléatoire	Agglomérée	Agglomérée	Agglomérée	Aléatoire
16	Agglomérée	Agglomérée	Agglomérée	Agglomérée	Aléatoire

Tableau 1.A.2 – Pourcentage de tests refusés de Kolmogorov-Smirnov à 10%

	Secteur A	Secteur B	Secteur C	Secteur D	Secteur E	Moyenne
Gini	13	19	4	48	2	17
WGini	25	17	21	21	10	19
Herfindahl	19	17	17	63	15	26
EG	29	21	50	10	23	27
MS	15	10	15	4	13	11
Moyenne	20	17	21	29	13	20

Tableau 1.A.3 – Pourcentage de tests refusés de Kolmogorov-Smirnov à 1%

	Secteur A	Secteur B	Secteur C	Secteur D	Secteur E	Moyenne
Gini	0	8	2	21	0	6
WGini	8	10	6	4	4	7
Herfindahl	6	2	4	38	0	10
EG	13	6	25	0	0	9
MS	8	4	2	0	0	3
Moyenne	7	6	8	13	1	7

CHAPITRE 2

Apports et limites des indices de concentration continus pour mesurer la concentration géographique des activités

Résumé

Par nature l'espace est continu mais sa discrétisation, du fait du regroupement spatial d'observations à une échelle géographique particulière, peut biaiser la valeur de certains indices (Herfindahl et Ellison-Glaeser par exemple). Le développement d'indices de concentration continus (Duranton et Overman, 2005; Marcon et Puech, 2010) permet de s'affranchir du choix d'un découpage géographique particulier, puisque ces indices de troisième génération se basent sur la prise en compte explicite de la distance pour apprécier la proximité géographique de deux localisations. Ce travail se propose d'identifier les situations dans lesquelles les indices continus peuvent avoir une portée limitée. Nous montrons que les valeurs de concentration obtenues à partir de ces indices peuvent être contradictoires et qu'ils soulèvent des difficultés quant au choix de la distance appropriée. De plus, nous verrons que les résultats peuvent être faussés en raison du niveau d'agrégation sectoriel choisi. De fait, l'utilisation de ces indices n'est pertinente que dans le cadre des contextes spécifiques et pour un seul secteur, plutôt que d'être utilisé dans le cadre d'études systématiques et exhaustives.

2.1 Introduction

L'étude des bénéfices de la proximité géographique des activités fait l'objet d'une littérature riche, aussi bien d'un point de vue théorique (Fujita *et al.*, 1999; Fujita et Thisse, 2002) qu'empirique (Head *et al.*, 1995). Selon Marshall (1890) les économies liées à l'agglomération spatiale sont de trois sortes : réduction des coûts de transport entre clients et fournisseurs, disponibilité d'une main d'œuvre spécialisée et stable et enfin, meilleure diffusion des connaissances (externalités technologiques).

La concentration géographique des activités peut aussi bien concerner des relations entre entreprises d'un même secteur, que des entreprises appartenant à des branches différentes.

Pour mesurer cette concentration spatiale, on utilise des indices de concentration si les établissements appartiennent à un même secteur et des indices de coagglomération si les établissements appartiennent à deux secteurs différents. La qualité d'un indice de concentration repose sur le respect d'un certain nombre d'hypothèses (Combes et Overman, 2004; Duranton et Overman, 2005) ; un indice doit être comparable entre secteurs, permettre de mesurer significativement les écarts entre zones, périodes ou secteurs, prendre en compte l'agglomération géographique de l'activité, être insensible à la classification sectorielle et au découpage géographique. Dans le cas où cette dernière propriété n'est pas respectée, il n'est pas possible de comparer des résultats obtenus à partir de découpages géographiques différents. On parle du problème du Modifiable Areal Unit Problem (MAUP). Mis en évidence par Openshaw et Taylor (1979, 1981), celui-ci est associé à deux effets. Un effet d'échelle si la variation de la mesure est due à l'agrégation (ou la désagrégation) d'unités géographiques et un effet de zonage si la variation est due à une modification des frontières sans que la taille et le nombre des zones ne soient modifiés.

Dans un premier temps, des indices conçus dans le contexte d'autres domaines ont été mobilisés pour étudier les phénomènes spatiaux (Gini, 1912; Herfindahl, 1950). Puis des indices spécifiquement dédiés à l'analyse de la concentration géographique des activités ont été développés (Ellison et Glaeser, 1997; Maurel et Sédillot, 1999). Ces indices de deuxième génération présentent néanmoins l'inconvénient de réduire l'espace à une somme d'entités géographiques issues d'un découpage spécifique (commune, département, région).

Les indices discrets obtenus à partir d'une discrétisation de l'espace selon un découpage géographique (la plupart du temps administratif) sont de fait sensible au MAUP. Des travaux montrent que ces indices discrets peuvent être biaisés à la hausse comme l'indice d'Ellison et Glaeser (Rosenthal et Strange, 2001), ou à la baisse comme l'indice de Gini (Briant *et al.*, 2010). L'indice discret

de Maurel et Sédillot (1999) est en revanche peu sensible à l'effet d'agrégation du MAUP.

De nouvelles mesures ont été développées et s'affranchissent théoriquement des problèmes liés à la discrétisation de la géographie des activités. Les indices de troisième génération de Duranton et Overman (2005) et Marcon et Puech (2003) sont souvent cités dans la littérature économique car régulièrement appliqués pour étudier la logique de localisation des activités sur un territoire. Ces mesures sont dites "continues" car la géographie est prise en compte par l'intermédiaire des distances qui séparent les établissements plutôt que leur appartenance à une entité administrative. De ce fait, ces mesures ne sont plus sensibles au MAUP. D'autres mesures continues peuvent être utilisées mais n'ont pas pour vocation l'étude de la concentration géographique des activités comme le K de Ripley (1976, 1977), D de Diggle et Chetwynd (1991) [1].

Comme dans le cas des indices discrets, de nombreux travaux portent sur le calcul de ces indices à partir de données empiriques. Behrens et Bougna (2015) étudient la concentration des activités canadiennes sur plusieurs années (2001, 2005 et 2009). Ils appliquent à la fois l'indice discret d'Ellison et Glaeser et l'indice continu de Duranton et Overman afin de comparer la sensibilité des résultats à l'approche discrète ou continue. L'apport de ces mesures continues est d'autant plus utile pour des études sur la répartition canadienne des activités que 90% des établissements sont à moins de 161 kilomètres (100 miles) de la frontière avec les États-Unis. La répartition des activités correspond à une bande d'activité frontalière plutôt qu'à une répartition relativement homogène comme pour le cas français ou britannique. Notons également la contribution plus spécifique de Barlet et Collin (2010) qui étudient la répartition des activités de santé en France même si le calcul de l'indice de Marcon et Puech (2010) se limite uniquement aux distances de 1 et 10 kilomètres et ne distingue pas les valeurs significatives des valeurs non-significatives.

L'apport des mesures continues par rapport aux mesures discrètes est important car il n'existe plus de choix *ex ante* pour favoriser un découpage géographique particulier. Notre travail se propose d'en étudier les limites possibles de manière plus appliquée que Marcon et Puech (2014). À partir de données simulées, nous examinons si ces indices continus, insensibles au MAUP, sont malgré tout sensibles à la position des établissements relativement aux limites du territoire de référence. Nous vérifions également les implications des paramètres choisis lors du calcul de ces indices continus, plus particulièrement le choix de la distance seuil jusqu'à laquelle la proximité géographique est imputable à la concentration géographique des établissements et donc, aux conséquences directes ou indirectes de liens entre les établissements du secteur. Enfin, nous discutons du problème de classification sectorielle en utilisant des indices discrets qui permettent de mettre en évidence les localisations différentes, ou non, d'établissements de même secteur à des niveaux plus agrégés. Le

1. Voir Marcon et Puech (2014) pour une revue détaillée des mesures.

problème de classification sectorielle est similaire à celui du MAUP, on parle du MSUP [2] mentionné par (Abdelmalki *et al.*, 2012; Puech, 2003). À la différence du MAUP, le MSUP est exlusivement lié à un effet d'échelle. Comme dans le cas de l'agrégation géographique, l'agrégation sectorielle peut biaiser les résultats. Fratesi (2008) illustre le problème du MSUP et du MAUP à l'aide d'une répartition des activités particulières. Avec cette répartition, il montre que l'agrégation sectorielle peut conduire à une perte d'information si les logiques de localisation des sous-secteurs sont différentes.

Nous présentons plus en détails les deux mesures retenues dans une deuxième section. Puis, nous discutons de leurs limites du fait de certains choix à partir de données empiriques dans une troisième section. Dans une quatrième section nous abordons plus en profondeur le problème de la classification sectorielle. Nous concluons ensuite sur les apports et limites de ces mesures continues en proposant des recommandations quant à l'utilisation de ces mesures.

2.2 Indices en continu

Nous distinguons les deux types de mesures énoncés par Haaland *et al.* (1999), les mesures *absolues* Duranton et Overman (2005) [3] et *relatives* Marcon et Puech (2003) [4]. Les mesures absolues sont les plus souvent utilisées et permettent de caractériser la concentration d'un secteur sans valeur de référence. Moins courantes, les mesures relatives permettent de comparer la concentration d'un secteur relativement au poids du secteur dans l'activité globale [5]. Dans les deux cas la significativité de ces mesures est évaluée en comparant la localisation effective à des localisations contrefactuelles. Dans ce qui suit nous présentons le détail de calcul de ces deux mesures et la manière dont sont obtenues ces localisations contrefactuelles. Notons également que les deux indices présentés différent aussi dans la manière de prendre en compte le voisinage. L'indice DO considère deux observations comme voisines si elles se situent à une distance précise notée d. L'indice M considère deux observations comme voisines si elles se situent dans un cercle dont le rayon est inférieur à la distance notée r.

2.2.1 Duranton et Overman

Pour évaluer si un secteur est concentré, Duranton et Overman (2005) évaluent la densité des établissements d'un secteur à chaque distance possible sur le territoire de référence. Pour une

2. Modifiable Sector-based Unit Problem.
3. DO par la suite.
4. M par la suite.
5. Il existe également des mesures topographiques (Brülhart et Traeger, 2005) ; pour ces mesures on prend en référence l'espace physique. Le nombre d'observations est ramené à l'aire du territoire. Comme ces mesures font l'hypothèse implicite que la répartition globale des activités est homogène, nous n'étudions pas ce type de mesure dans le chapitre.

industrie donnée avec n établissements, on calcule les $\frac{n(n-1)}{2}$ distances euclidiennes entre les établissements. Soit $d_{i,j}$ la distance entre les établissements i et j. La densité à chaque point d est estimée en considérant un noyau de fenêtre h à partir d'une densité f gaussienne :

$$\hat{K}(d) = \frac{1}{n(n-1)h} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} f\left(\frac{d - d_{i,j}}{h}\right)$$

Une valeur est ainsi obtenue pour chaque distance d, mais elle n'est pas directement interprétable. Il est nécessaire de pouvoir la comparer à un ensemble d'autres valeurs qui justifie la simulation de localisations contrefactuelles à partir desquelles sera établi un intervalle de confiance. On pourra conclure à de la concentration ou à de la dispersion des établissements du secteur lorsque la mesure calculée se situe en dehors des bornes de cet intervalle.

Pour étudier les localisations d'un secteur industriel particulier sur un territoire national, il s'agit traditionnellement de considérer l'ensemble des implantations de tout le secteur industriel sur tout le territoire comme des implantations qui auraient pu être choisies par les établissements du secteur spécifique. Puis on procède à de nouveaux réarrangements de ces implantations, c'est-à-dire que l'on relocalise aléatoirement les établissements parmi les localisations possibles. Chacun de ces réarrangements constituant une simulation contrefactuelle, mais possible, de la localisations des établissements du secteur. On peut alors conclure ou non quant à la concentration ou dispersion géographique en comparant la localisation observée à un intervalle de confiance calculé à partir des localisations simulées.

2.2.2 Intervalles de confiances

De manière simple, les localisations possibles caractérisent les localisations des établissements qui partagent des caractéristiques communes : appartenance au même secteur d'activité, établissements de même taille, etc. Ces localisations possibles ne peuvent pas être choisies au-delà d'une distance seuil (la distance médiane entre deux établissements sur l'ensemble du territoire par exemple). Le choix de ne pas étudier les localisations au-delà de cette distance repose sur l'hypothèse que pour ces établissements, la localisation ne constitue pas un facteur de productivité significatif.

La comparaison se fait ensuite entre la distribution effective des localisations et les distributions des localisations contrefactuelles. L'intervalle de confiance obtenu à partir des simulations contrefactuellles peut être local ou global. Les différences entre ces deux intervalles sont présentées en reprenant les notations de l'indice de Duranton et Overman (2005) [6].

6. Les intervalles de confiance peuvent être obtenus de la même manière pour l'indice Marcon Puech.

Intervalle de confiance local

Pour chaque valeur de l'intervalle des distances (de 0 kilomètre jusqu'à la distance maximum $\overline{d}$ selon le critère retenu), on ordonne les valeurs des 1000 simulations contrefactuelles. Pour un niveau de confiance de 95% [7], la valeur du 95ème centile de cette distribution constituera la valeur de la bande haute de l'intervalle de confiance : $\overline{K}_A(d)$. S'il existe une distance d telle que $\hat{K}_A(d) > \overline{K}_A(d)$ on peut conclure à de la concentration géographique des établissements du secteur à la distance d. De la même façon, la valeur du 5ème centile déterminera la valeur de la bande basse de l'intervalle de confiance $\underline{K}_A(d)$. S'il existe une distance d telle que $\hat{K}_A(d) < \underline{K}_A(d)$ on peut conclure à de la dispersion à la distance d .

L'indice de concentration à la distance d pour un secteur A est donc :

$$\gamma_A(d) = \max\left(\hat{K}_A(d) - \overline{K}_A(d), 0\right) \tag{2.1}$$

L'écart maximum constaté sur $[0\,;\overline{d}]$, entre la la localisation effective et la bande haute de l'intervalle de confiance).

Par symétrie, l'indice de dispersion à la distance d est :

$$\psi_A(d) = \max\left(\underline{K}_A(d) - \hat{K}_A(d), 0\right) \tag{2.2}$$

soit l'écart maximum constaté sur $[0\,;\overline{d}]$, entre la bande basse de l'intervalle de confiance et la localisation effective.

Même si ce mode de détermination est relativement simple, on supprime des valeurs à une distance donnée tout en gardant le reste de la distribution simulée sans prendre en compte l'autocorrélation des valeurs d'une même distribution. Or, supposons qu'une simulation soit aberrante sur certaines distances compte tenu des 999 autres, on considérerait donc la partie aberrante comme concentrée (ou dispersée) tandis que le reste de la simulation ne serait pas considérée comme faisant partie de cette simulation concentrée. En médecine, cela équivaudrait à considérer qu'un traitement est plus efficace s'il soigne plus rapidement, sans vérifier si ce même traitement augmente les risques de rechute sur le long terme. Les conclusions seraient donc fallacieuses.

Intervalle de confiance global

Pour déterminer l'intervalle de confiance global, on ne considère plus les valeurs de toutes les simulations de localisations à une distance d, mais chaque localisation contrefactuelle comme une

7. Dans ce chapitre, tous les intervalles de confiance sont déterminés pour un seuil de 5%.

observation pour l'ensemble des distances d. Toujours pour un niveau de confiance de 95%, et à partir de 1000 simulations sur l'intervalle $[0\,;\overline{d}]$, on souhaite disposer d'une bande haute ($\overline{\overline{K}}_A$) telle que $5\% \times 1000 = 50$ simulations montrent au moins une fois (et ce quelle que soit la distance) de la concentration. De manière symétrique, la bande basse ($\underline{\underline{K}}_A$) est telle que 50 simulations amènent à conclure à de la dispersion au moins une fois quelle que soit la distance. Notons dans ce cas que l'on utilise les 950 simulations ne montrant pas de concentration car un secteur ne peut pas être simultanément concentré et dispersé. Le secteur est dit concentré si la valeur est au moins une fois supérieure à la bande haute de l'intervalle de confiance, la valeur retenue est l'écart maximum constaté pour $d \in [0; \overline{d}]$:

$$\Gamma_A(d) = max(\hat{K}_A(d) - \overline{\overline{K}}_A(d), 0)$$

Le secteur est dit dispersé si le secteur n'est pas concentré et si la valeur est au moins une fois inférieure à la bande basse, la valeur retenue est l'écart maximum constaté pour $d \in [0; \overline{d}]$:

$$\Phi_A(d) = \begin{cases} max\left(\underline{\underline{K}}_A(d) - \hat{K}_A(d), 0\right) \text{ si } \Sigma_{d=0}^{d=\overline{d}}\, \Gamma_A(d) = 0, \\ 0 \text{ sinon,} \end{cases}$$

Dans les faits, les bornes des intervalles changent peu que l'intervalle de confiance soit local ou global.

2.2.3 Marcon et Puech

La mesure proposée par Marcon et Puech (2010) repose sur une approche relative : pour une distance maximum r est-ce qu'il y a plus de chances de trouver des entreprises similaires relativement à l'ensemble des autres entreprises ?

Pour chaque établissement localisé en un point et pour une distance donnée r, on note T_r le nombre d'établissement du même secteur situés à une distance inférieure à r et N_r le nombre d'établissements de tous les secteurs. Pour chaque valeur de r on calcule un ratio de voisins cibles T_r/N_r ce qui permet de calculer pour l'ensemble des établissements un ratio moyen noté $\overline{T_r/N_r}$. Par ailleurs sur le territoire, le ratio global est égal à T/N.

Si $\overline{T_r/N_r} > T/N$ pour une distance r, cela signifie qu'il y a proportionnellement plus d'établissements du même secteur à une distance r que sur l'ensemble du territoire [8]. On désigne par M

8. La pondération par le nombre d'employés est également possible mais dans le chapitre nous n'étudierons pas l'impact qu'aurait des tailles d'établissements différentes. Pour reprendre la formulation de Giuliani *et al.* (2014), nous étudions la concentration géographiques des firmes (sans pondération des points) et non la concentration géographique des activités économiques (si l'on pondère les points par la taille des emplois).

l'indice de Marcon et Puech :

$$M(r) = \frac{\overline{T_r/N_r}}{T/N} \tag{2.3}$$

Si l'on désire étudier un seul secteur et que $M(r) = z$, jusqu'à la distance r, la densité relative des établissements du secteur est z fois plus importante que celle du territoire global. Plus r est élevé, plus grand sera le territoire, donc plus le ratio tendra vers 1. Cet indice présente l'avantage d'être directement interprétable même si la valeur est parfois utilisée sans que sa significativité ne soit établie. Les intervalles de confiance sont ensuite obtenus de la même manière que pour l'indice DO. Il est toujours nécessaire d'effectuer un choix quant au localisations possibles avant d'effectuer les simulations de localisations contrefactuelles.

2.2.4 Exemples

Pour illustrer l'utilisation de ces indices, nous répliquons les distributions proposées par Marcon et Puech (2010). Le Tableau 2.1 présente les répartitions d'établissements pour les trois exemples [9].

Tableau 2.1 – Répartitions des établissements des trois illustrations

	Exemple 1	Exemple 2	Exemple 3
Secteur d'intérêt	Poisson 50 établissements	Matérn (cluster) 50 établissements	9 clusters 50 établissements
Localisations contrefactuelles	Poisson 500 localisations	Poisson 200 localisations	Poisson 500 localisations

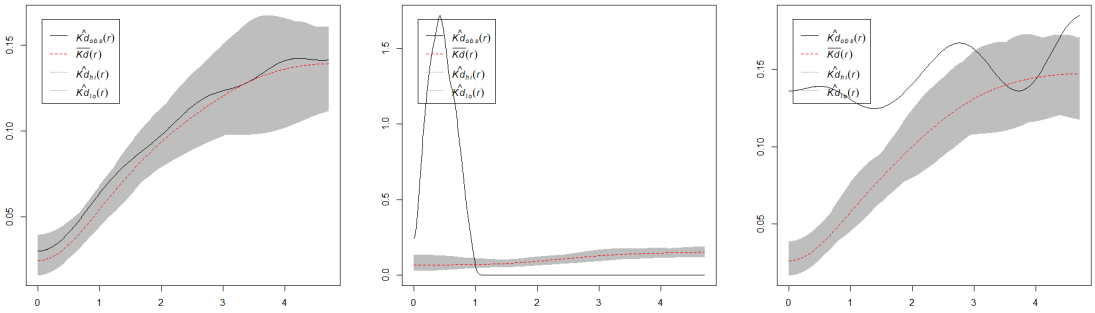
Pour l'exemple 1, aucun des indices ne conclut à la concentration ou à la dispersion des établissements car la fonction représentative de chacun des indices (courbe en trait plein) est à l'intérieur de l'intervalle de confiance (zone grise). Ceci est attendu car les localisations des établissements du secteur d'intérêt et les localisations contrefactuelles sont issues d'un même processus de Poisson, les différences ne portant que sur le nombre de points considérés.

Pour l'exemple 2, les indices sont concordants pour conclure à la concentration des établissements. Néanmoins le caractère absolu de l'indice DO indique que la concentration se fait à une échelle inférieure à une distance de 1, alors que l'indice M indique qu'il y a concentration jusqu'à une distance de 7. Cette différence tient au fait que l'indice M prend en compte les établissements à une distance inférieure à une valeur donnée alors que l'indice DO prend en compte les établissements

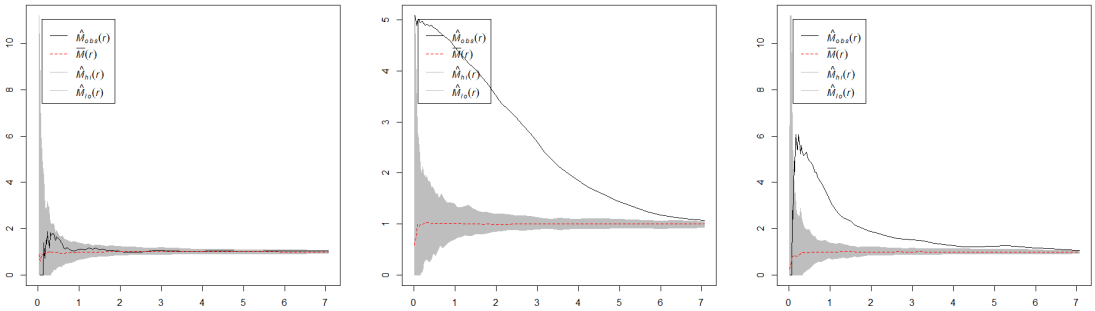
9. Dans la partie empirique de ce chapitre, nous désignons par cluster une agglomération de points similaires sans considérer que cela soit le résultat de choix particuliers, à la différence de Porter (2000) qui rattache la notion de cluster aux connexions et à la dépendance des établissements faisant partie de ce cluster. Ce chapitre étant dédié à la pertinence des mesures continues pour révéler la proximité géographique des établissements, nous n'avons pas besoin de savoir l'intensité des liens et échanges entre établissements.

La distance seuil prise par défaut est de 4.71, soit le tiers de la distance maximale possible, comme décrit par Duranton et Overman (2005). Le rayon utilisé pour le processus de Matérn est de 0.5. L'intensité du processus de Poisson est telle que l'on obtient en espérance le nombre d'établissements désirés sur le territoire carré de longueur 10. C'est-à-dire un paramètre de 2 pour implanter 200 établissements, 5 pour 500 établissements,...

FIGURE 2.1 – Fonctions représentatives des indices continus



(a) Indice de Duranton et Overman pour les exemples 1 à 3



(b) Indice Marcon et Puech pour les exemples 1 à 3

à une distance précise.

L'exemple 3 permet de mettre en lumière la difficulté de l'indice M de donner une information pertinente quand il y a plusieurs clusters pour un secteur. En effet, il y a de la concentration à chaque distance comme pour le deuxième exemple. L'indice DO en revanche permet de saisir la multipolarité de la répartition car il y a de la concentration jusqu'à une distance de 3 puis à partir d'une distance de 4.

Ainsi, comme ces exemples permettent de le montrer, on observe des différences entre ces deux indices et on met ainsi en évidence les conséquences du caractère absolu ou relatif des mesures prises en compte, notamment pour l'exemple 3. Les approches relatives, comme celle de l'indice M, ne permettent pas de différencier une répartition en un pôle d'agrégation ou en plusieurs clusters. Dans la prochaine section nous allons détailler quelles peuvent être les autres limites qui contraignent l'utilisation des méthodes continues, aussi bien avant leur application que pendant la phase de calcul.

2.3 Limites des indices continus

2.3.1 Accès aux données

Le calcul des indices discrets repose sur des observations localisées selon une entité géographique (commune, département,...), tandis que pour calculer des indices continus il faut disposer

d'information sur leur géolocalisation la plus précise possible. Le respect des règles d'anonymat des données individuelles peut être un frein pour calculer ces indices. En effet, afin de préserver cet anonymat, les instituts statistiques en charge de la collecte puis de la diffusion des données agrègent celles-ci de sorte que pour chaque zone géographique considérée il y ait au minimum un certain nombre d'individus pour éviter le risque de divulgation (Buron et Fontaine, 2018; INSEE, 2010).

Ainsi, en dépit des facilités d'accès à des données de plus en plus exhaustives, il reste qu'il existe toujours un arbitrage à faire entre la précision géographique et la qualité de l'information disponible à partir de ces données ou que l'on peut restituer dans les résultats.

2.3.2 Choix de l'unité de distance

Pour calculer les indices continus, on utilise le plus souvent les distances à vol d'oiseau (distance euclidienne). D'autres distances peuvent être prises en compte. Par exemple, les distances kilométriques à partir du réseau routier pourraient apporter une information plus précise mais nécessitent plus de temps de calcul. De plus, Combes et Lafourcade (2005) montrent que dans le cas français la distance kilométrique et en coût de transport ont une corrélation de 0.97 avec la distance euclidienne, il n'y a donc que peu de différences entre la distance euclidienne et la distance en coût de transport.

Une autre alternative serait de considérer les localisations accessibles en dessous d'un temps prédéterminé. De plus, la correction de la distance euclidienne par la distance géodésique peut être souhaitable quand le territoire étudié est étendu.

En fait, on se retrouve dans une situation analogue à celle décrite dans le cas du choix des critères pour définir la matrice de pondération dans le cas d'unités spatiales discrètes. Dans certains cas, ce choix peut dépendre de la nature du problème considéré et poser des problèmes d'identification.

2.3.3 Présentation des résultats avec différents seuil de distance

Une difficulté majeure vient de la comparaison de travaux sur des territoires différents. En effet, si la concentration est analysée jusqu'à la distance de 800 kilomètres sur le territoire canadien (Behrens et Bougna, 2015), la distance retenue pour le Royaume-Uni est de 180 kilomètres (Duranton et Overman, 2005) et 392 kilomètres pour le cas français (Barlet *et al.*, 2008). Même si ce choix est justifié par la médiane des distances entre les établissements sur tout le territoire, cela veut aussi dire que selon le territoire étudié la distance et la proximité géographique ne sont plus appréhendées de la même façon par les individus et les établissements. Cela correspond, si l'on utilise des indices discrets, à utiliser des zones plus grandes si le territoire est plus grand. Les travaux sur données

japonaises de Nakajima *et al.* (2012) utilisent une distance seuil de 180 kilomètres comme pour le Royaume-Uni afin de permettre une meilleure comparaison, bien que la distance médiane soit d'environ 400 kilomètres. Kosfeld *et al.* (2011) utilisent une fonction semblable à celle de Besag (1977) sur données allemandes en posant la distance seuil comme le quart de la distance maximale, soit 215 kilomètres. La fonction de Ripley est privilégiée par Albert *et al.* (2012) en posant une distance limite de 200 kilomètres sur données espagnoles mais qui n'est justifiée par aucun critère objectif.

Les résultats pour des indices discrets peuvent être sensibles au choix du découpage géographique retenu. En revanche, le calcul d'indices continus supposent le choix d'une distance spécifique. Présenter les résultats de calculs d'indices continus pour l'ensemble des secteurs n'est donc pas chose aisée : les logiques de localisation sont très variées et les résultats peuvent aussi être présentés pour différentes distances (le Tableau 2.2 présente les résultats de Barlet *et al.* (2008) à partir de la mesure de Duranton et Overman (2005) pour la France métropolitaine). Le choix des distances utilisées (5, 30 et 150 km) pour décrire la concentration des secteurs est similaire au choix du découpage géographique pour les indices discrets et ne permet qu'une synthétisation difficile des résultats. Il n'est pas possible de résumer la concentration géographique d'un secteur à une variable binaire.

Tableau 2.2 – Part de secteurs manufacturiers concentrés à différentes distances d'après Barlet *et al.* (2008)

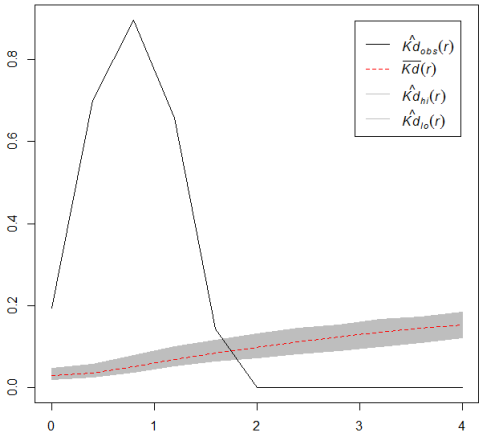
5km	5km seul	5 & 30 km seul	5 & 150 km seul	5 & 30 & 150 km
34 %	0 %	24 %	0 %	10 %
30km	30km seul	30 & 150 km seul		
38 %	2 %	2 %		
150km	150km seul			
22 %	10 %			

Pour les indices DO et M, nous utiliserons à la fois la valeur cumulée comme calculé par DO et la valeur maximum de la concentration [10]. Pour illustrer le problème du choix de la distance seuil, nous avons simulé des localisations pour deux industries. La première industrie comporte 50 établissements répartis selon un processus de Matérn au centre du territoire avec un radius de 1. Les 500 établissements du deuxième secteur sont répartis de manière aléatoire sur l'ensemble du territoire. Nous procédons aux calculs des indices DO et M en faisant varier la distance seuil de 1 à 14. La valeur suggérée par les auteurs aurait été de 4.71. La Figure 2.2 présente les graphiques pour des distances seuil de 4 puis 12 et la Figure 2.3 présente l'évolution de la valeur des indices selon la distance seuil.

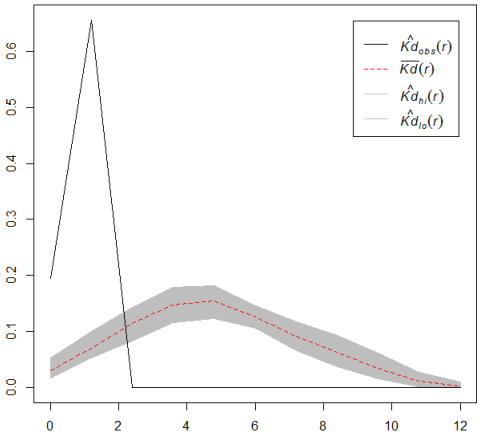
Les résultats sont attendus car en augmentant la distance seuil, la valeur maximum est plus

10. La valeur maximum est l'écart maximum entre la valeur observée et la bande haute de l'intervalle de confiance.

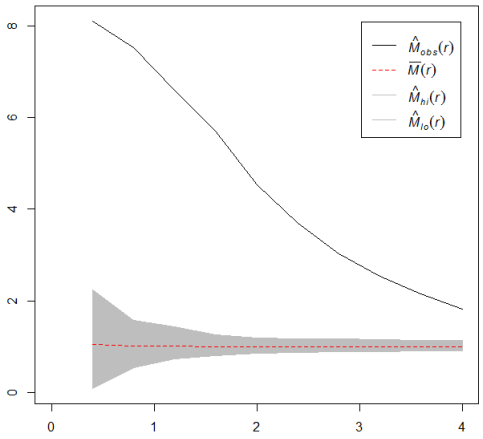
FIGURE 2.2 – Exemples de la variation de la distance seuil



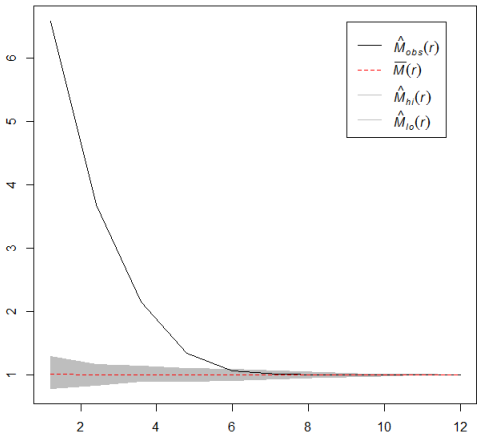
(a) Indice DO pour une distance de 4



(b) Indice DO pour une distance de 12

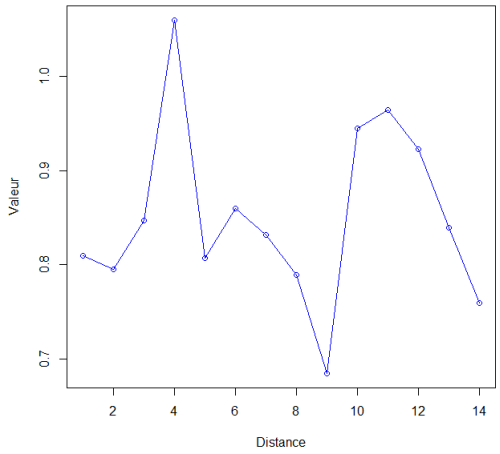


(c) Indice M pour une distance de 4

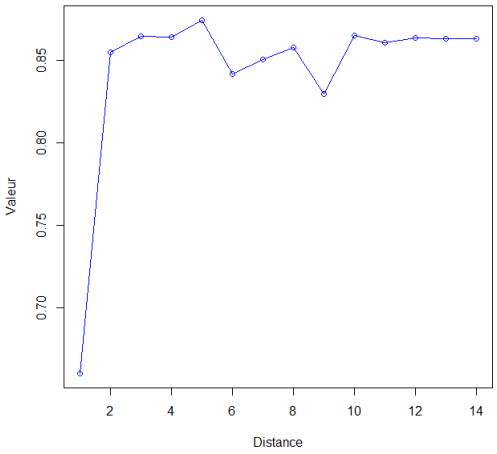


(d) Indice M pour une distance de 12

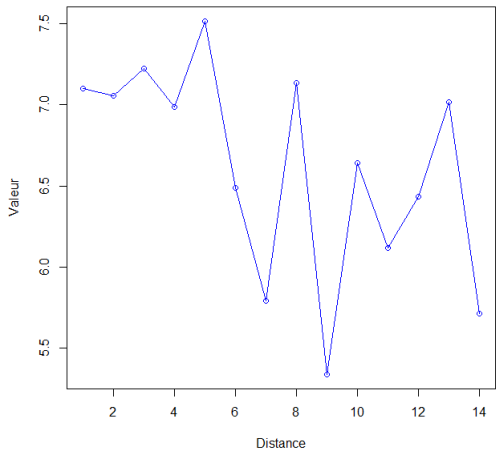
FIGURE 2.3 – Effets de la variation de la distance seuil



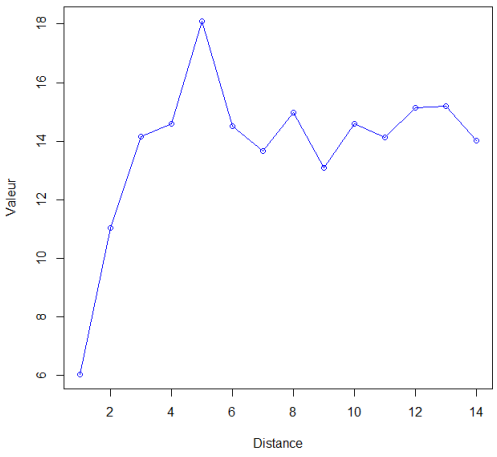
(a) Valeur maximale de l'indice DO



(b) Valeur cumulée de l'indice DO



(c) Valeur maximale de l'indice M



(d) Valeur cumulée de l'indice M

faible, l'écart cumulé est alors d'autant plus faible.

Notons que les mesures permettent de conclure à la concentration du secteur quels que soient la méthode ou l'indice utilisé, et jamais à la dispersion des établissements. Qualitativement, les mesures sont donc validées. On constate que la valeur cumulée de l'indice DO est plus stable que la valeur maximale, cette dernière oscillant entre 0.68 et 1.06 alors que la valeur cumulée est d'environ 0.85. Pour l'indice M, on constate un pic de concentration à une distance seuil de 5, avant qu'il ne diminue puis se stabilise autour de 1.4 pour la valeur cumulée ; la valeur maximale est variable et oscille entre 5.4 et 7.5. Nous pouvons néanmoins déduire que la prise en compte de la valeur cumulée est moins sensible à un changement de la distance. Autre fait intéressant, nous trouvons dans tous les cas que la concentration maximale détectée quand la distance correspond à la valeur seuil par défaut (entre 4 et 5). Dans ce cas, la valeur par défaut conduit à conclure à une concentration géographique plus forte que pour toute autre distance seuil. Sur données empiriques il serait intéressant de vérifier si ce choix par défaut de la distance seuil n'aurait pas une influence significative sur les résultats. Dans ce cas précis, les valeurs de l'indice DO varient moins que les valeurs de l'indice M, ce qui incite plutôt à favoriser l'utilisation de l'indice DO.

2.3.4 Variation de la position du cluster

La question est ici de savoir si la valeur d'un indice est identique si le cluster des établissements est au centre du territoire ou excentré. La sensibilité à la position relative des établissements est rarement mentionnée dans la littérature. En effet, il est habituellement considéré que l'on cherche à mesurer la concentration des établissements sur le territoire conditionnellement à leur localisation. Il n'est en revanche jamais question d'étudier si la position relative des établissements du territoire de référence peut influencer la valeur de l'indice, et si tel est le cas, si l'impact est positif (concentration) ou négatif (dispersion). Empiriquement, cela relève de la comparabilité de secteurs ne nécessitant pas des besoins localisés particuliers (ressources naturelles par exemple) à des secteurs ayant des besoins spécifiques (les activités portuaires). De plus, la position relative des établissements sur un territoire, et donc la proximité des frontières, est liée à la géographie du territoire dans son ensemble. Aussi, même si l'utilisation de la distance permet de ne pas dépendre d'un choix géographique particulier (et donc le MAUP), il est également nécessaire de vérifier si les indices ne sont pas dépendants de la position relative des établissements du secteur étudié.

Pour vérifier si la position relative d'un cluster sur le territoire modifie la valeur de la mesure de concentration, nous allons faire varier la position d'un cluster sur un territoire. Pour cela, nous supposons que l'ensemble des activités est réparti de manière homogène [11] et que les établissements

11. On procède à un tirage des coordonnées géographiques dans une loi uniforme.

d'un nouveau secteur s'implantent sous la forme d'un cluster. Le territoire est un carré de côté égal à 10 et les établissements du secteur étudié A sont dans un cluster dont les localisations sont obtenues à partir d'un processus de Matérn [12] et de rayon égal à 1. Toutes les autres activités sont réparties de manière aléatoire conduisant en espérance à une répartition homogène. Nous faisons varier le centre du cluster sur le territoire afin de vérifier si la position relative des établissements du secteur influence la valeur de la concentration. Quand le centre du secteur se rapproche de la frontière, il est possible que des établissements se retrouvent en dehors du territoire et il y aurait de moins en moins d'établissements du secteur d'intérêt. Dans ce cas, nous relocalisons les établissements extérieurs à l'intérieur du territoire par symétrie axiale en utilisant la frontière dépassée. Par exemple, si les coordonnées d'un établissement sont (-0.1,2), les coordonnées après correction sont (0.1,2). Ainsi les établissements du cluster sont dans un cercle. Quand le centre du cluster se rapproche beaucoup d'une frontière, les établissements sont dans un demi cercle. Enfin, quand le centre est proche de deux frontières (dans un coin du territoire) alors les établissements sont dans un quart de cercle. Il y a toujours 50 établissements dans le secteur concentré et il y a 500 autres établissements dont les localisations auraient pu être choisies.

Nous utilisons la distance seuil par défaut et les résultats sont synthétisés dans la Figure 2.4.

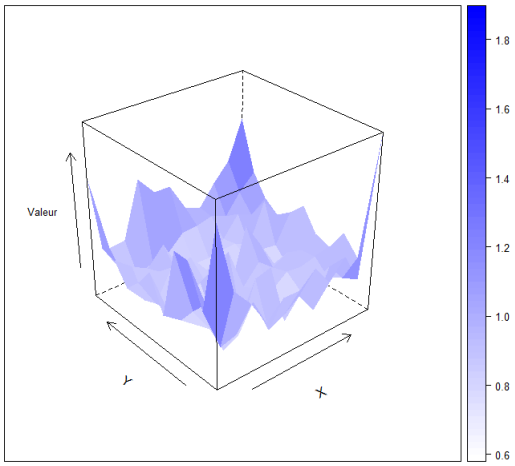
Notons tout d'abord qu'en aucun cas le secteur est dispersé, mais la position relative du cluster impacte la valeur mesurée de la concentration. Globalement nous constatons une augmentation de la concentration (valeur maximale et cumulée) quand l'épicentre sectoriel est proche d'une frontière et la concentration est encore plus élevée quand l'épicentre est dans un coin du territoire.

Comme précédemment, nous examinons dans quelle mesure le choix de la distance peut faire varier ces conclusions. En posant que la distance seuil est la distance maximale possible (la diagonale du territoire), nous effectuons les mêmes déplacements de l'épicentre sectoriel (Figure 2.5). Les graphiques montrent des résultats similaires pour l'indice M, à savoir une augmentation de la concentration quand l'épicentre est proche des frontières. Néanmoins, pour l'indice DO, l'augmentation de la distance conduit non plus à l'augmentation de la concentration quand l'épicentre est proche de la frontière, mais à la diminution de la concentration mesurée. Encore une fois, le choix de la distance n'est pas anodin, et particulièrement pour l'indice DO. En conséquence, les activités frontalières (comme les activités maritimes ou montagnardes par exemple), seront d'autant plus concentrées si l'on privilégie l'indice M que l'indice DO.

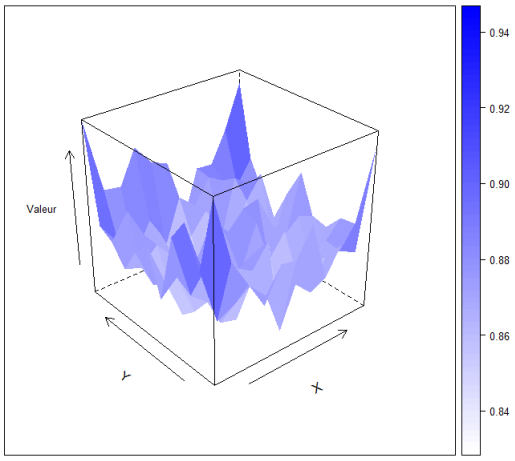
Nous avons vérifié si cela était lié à la relocalisation des établissements par symétrie axiale car notre méthode de relocalisation maintient le nombre d'établissements sur le territoire mais diminue la zone où ces établissements se localisent. Nous trouvons deux résultats différents selon l'indice

12. L'intensité du processus de Matérn permet d'avoir en moyenne le nombre d'établissements voulu.

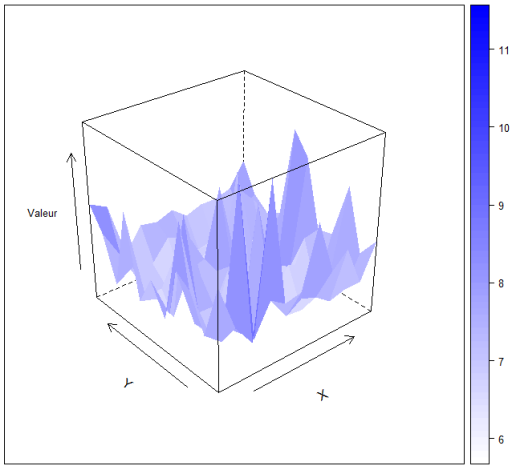
FIGURE 2.4 – Effets de la variation de la position du cluster avec la distance seuil par défaut



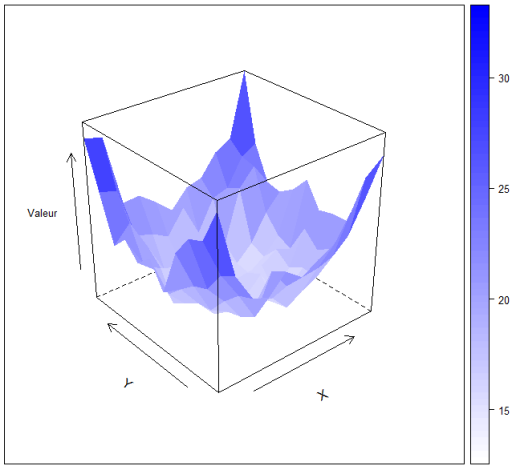
(a) Valeur maximale de l'indice DO



(b) Valeur cumulée de l'indice DO

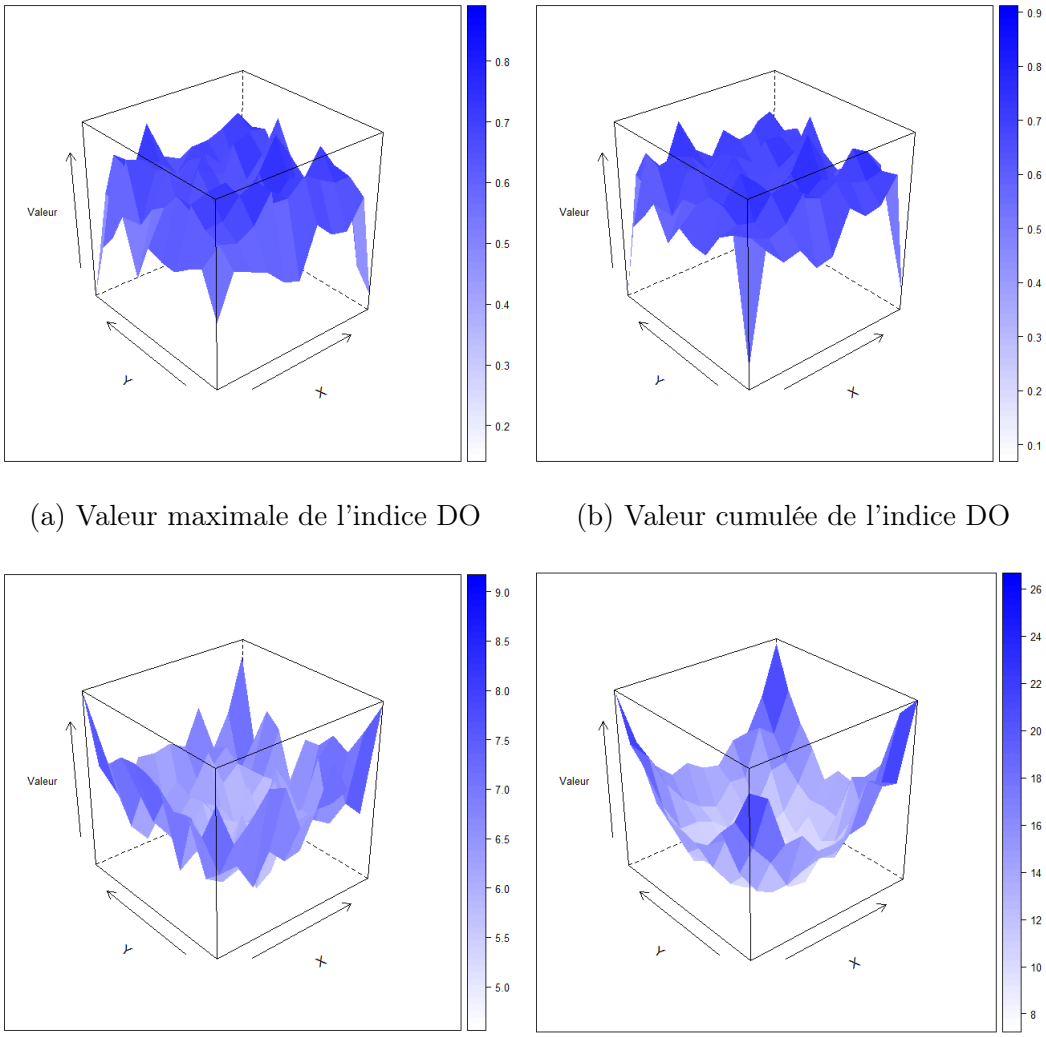


(c) Valeur maximale de l'indice M



(c) Valeur cumulée de l'indice M

FIGURE 2.5 – Effets de la variation de la position du cluster avec la distance seuil maximale



(a) Valeur maximale de l'indice DO

(b) Valeur cumulée de l'indice DO

(c) Valeur maximale de l'indice M

(d) Valeur cumulée de l'indice M

utilisé. Pour l'indice DO, les résultats sont très proches. Ce qui suggère que l'indice est peu sensible à la variation du nombre d'établissements. En revanche, pour l'indice M, les valeurs obtenues sont d'autant plus élevées lorsque les établissements restants sont proches des frontières. L'indice M est donc plus sensible au nombre d'établissements que l'indice DO.

Pour les indices continus, moins il y a d'établissements proches des frontières plus la concentration est significativement élevée. Ce résultat est en contradiction avec ce que l'on trouve à partir des indices discrets (Ellison et Glaeser, 1997; Maurel et Sédillot, 1999) car plus il y a d'établissements d'un secteur dans une même aire géographique, plus la concentration devrait être élevée. De même, moins il y a d'établissements d'un secteur, plus il y a de chance que la concentration observée soit le fruit du hasard, et donc issue d'une répartition aléatoire.

Nous avons vérifié la corrélation (au sens de Spearman et de Pearson) du nombre d'établissements avec la valeur des indices (Tableau 2.3). Quel que soit l'indice, il y a une corrélation négative entre le nombre d'établissements et la valeur de l'indice. Néanmoins, cette corrélation pourrait aussi être due au fait que les établissements sont peu nombreux quand l'épicentre sectoriel est proche des frontières, donc loin du centre du territoire.

Tableau 2.3 – Corrélation avec le nombre d'établissements

	Mmax	Mcum	KdMax	KdCum
Pearson nombre d'établissements variable	-0.82	-0.92	-0.65	-0.26
Spearman nombre d'établissements variable	-0.83	-0.83	-0.54	-0.34

Les corrélations sont significatives au seuil de 1%.

Nous avons également vérifié les corrélations entre la valeur des indices et la distance au centre du territoire (Tableau 2.4). Les corrélations confirment que la position relative a un impact sur la valeur des indices de concentrations continus. Si pour l'indice M la distance par rapport au centre augmente la concentration, celle-ci diminue pour l'indice DO. De plus, la modification de l'ajustement des localisations change l'impact de la distance au centre pour l'indice DO. Si quand le nombre d'établissements est fixe la valeur de l'indice tend à diminuer quand la distance augmente, à l'inverse quand les établissements sont moins nombreux près des frontières la valeur augmente avec la distance.

Pour l'indice DO et pour un même nombre d'établissements, la diminution de la concentration est cohérente avec le fait de trouver des valeurs de concentration plus faibles (ou même de la dispersion) pour des activités frontalières (Barlet *et al.*, 2008). Le fait de diminuer le nombre d'entreprises modifie le signe de la corrélation car si la distance augmente mais que le nombre d'établissements diminue, alors l'effet de la diminution du nombre d'établissements domine l'effet

de l'augmentation de la distance par rapport au centre. Pour l'indice M en revanche, l'augmentation de la distance du cluster par rapport au centre augmente toujours la valeur de la concentration.

Non seulement la position relative du cluster sur le territoire peut impacter la concentration mesurée, mais il possible que le nombre d'établissements puisse jouer.

Tableau 2.4 – Corrélation entre la valeur de l'indice et la distance au centre

	Mmax	Mcum	KdMax	KdCum
Pearson 50 établissements	0.64	0.91	-0.55	-0.50
Spearman 50 établissements	0.67	0.96	-0.53	-0.46
Pearson nombre d'établissements variable	0.47	0.73	0.44	0.32
Spearman nombre d'établissements variable	0.57	0.90	0.41	0.35

Les corrélations sont significatives au seuil de 1%.

Nous pouvons donc en conclure que les indices surestiment la concentration d'un secteur quand les établissements du secteur s'agglomèrent près des frontières.

2.3.5 Variation du nombre d'établissements

Quand on utilise des indices discrets pour mesurer la concentration, le nombre d'établissements du secteur considéré a un impact : moins il y a d'établissements, plus la concentration est élevée. Nous souhaitons vérifier si ce résultat se retrouve dans le cas des indices continus, avec la difficulté que ceux-ci considèrent le nombre d'établissements du secteur d'intérêt mais également le nombre de localisations qui auraient pu être choisies (les localisations contrefactuelles).

L'identification du biais potentiel lié au nombre d'établissements peut être influencée par la logique de simulation des localisations retenues. Nous considérons donc à trois logiques de simulation des localisations sur un territoire carré de côté égal à 10. Nous simulons les localisations pour deux secteurs (A et B) et mesurons à chaque fois la concentration d'un secteur en considérant les localisations de l'autre secteur comme contrefactuelles. Pour la première simulation, les établissements des deux secteurs sont répartis de manière aléatoire ce qui conduit à une répartition homogène sur tout le territoire. Pour la deuxième simulation, le secteur A est réparti de manière homogène et le secteur B selon un processus de Poisson. Pour la troisième simulation, les deux secteurs sont répartis selon un processus de Poisson.

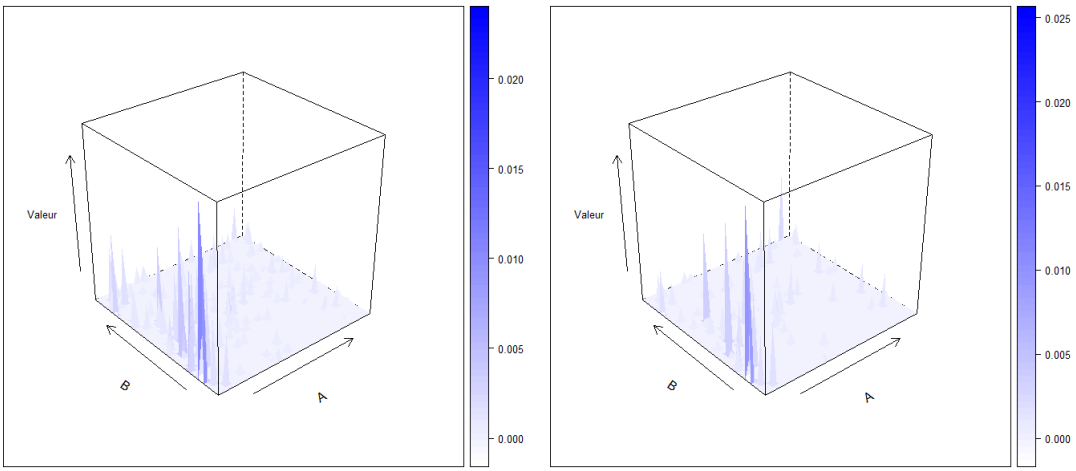
Pour chacune de ces simulations, le nombre d'établissements de chaque secteur varie de 10 à 500 avec un pas de 10.

Dans certains cas l'étude de la concentration se fait donc pour des secteurs avec beaucoup d'établissements avec des localisations contrefactuelles peu nombreuses. Dans de telles configurations, les simulations contrefactuelles seront très similaires et il y aura peu d'écart entre les bandes hautes

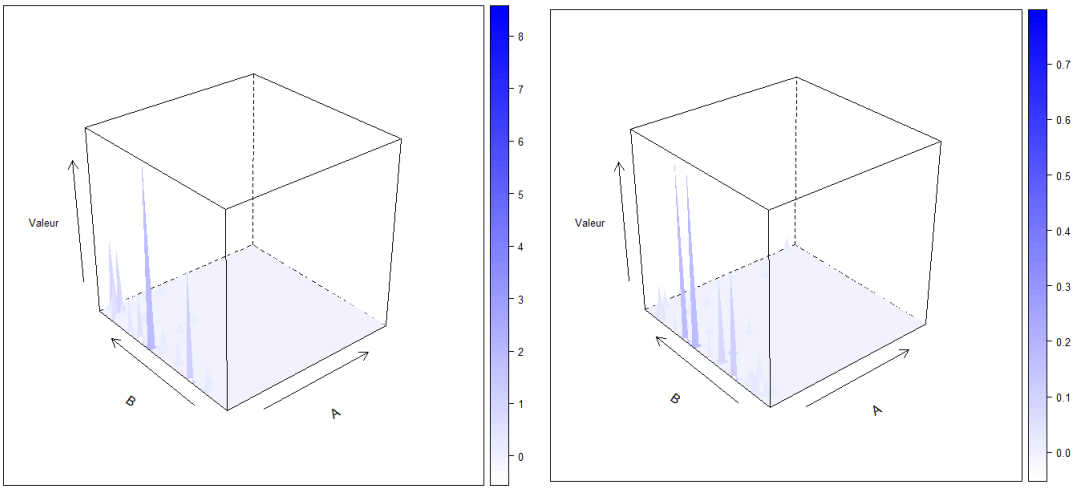
et basses de l'intervalle de confiance. Les Figures 2.6 à 2.9 illustrent les résultats pour les trois types de simulations.

Première logique de simulation

FIGURE 2.6 – Première logique de localisation : Secteur A



(a) Valeur maximale de l'indice DO (b) Valeur cumulée de l'indice DO

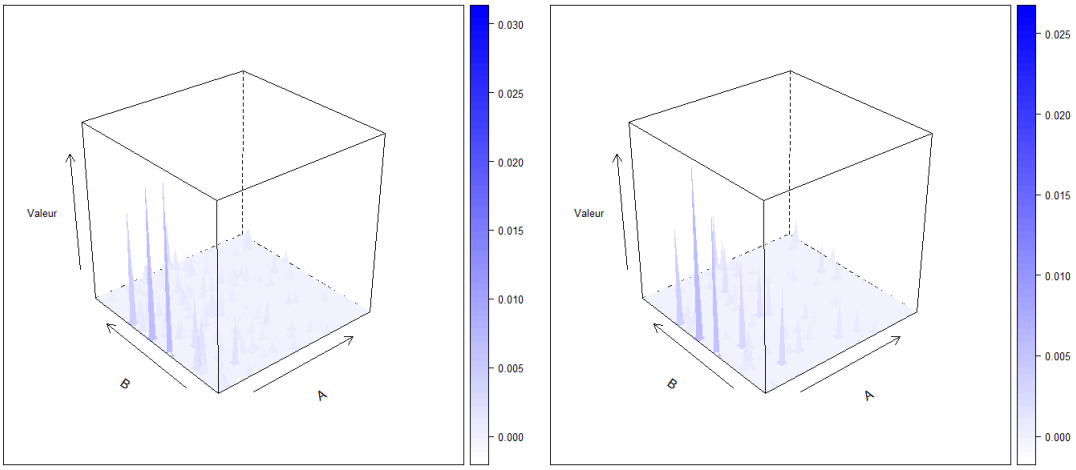


(c) Valeur maximale de l'indice M (d) Valeur cumulée de l'indice M

Comme les logiques de localisations sont similaires pour les deux secteurs, il n'est pas utile de présenter les résultats détaillés pour chacun d'entre eux : nous ne présentons donc que les résultats pour le secteur A. On constate que l'on peut généralement conclure à la concentration des établissements du secteur quand celui-ci comporte peu d'établissements. Ce résultat est vérifié pour les deux mesures même si l'indice M est plus sensible quand le nombre d'établissements est très faible. L'indice DO conduit également à conclure à la concentration des activités même si les pics de concentration sont relativement moins élevés.

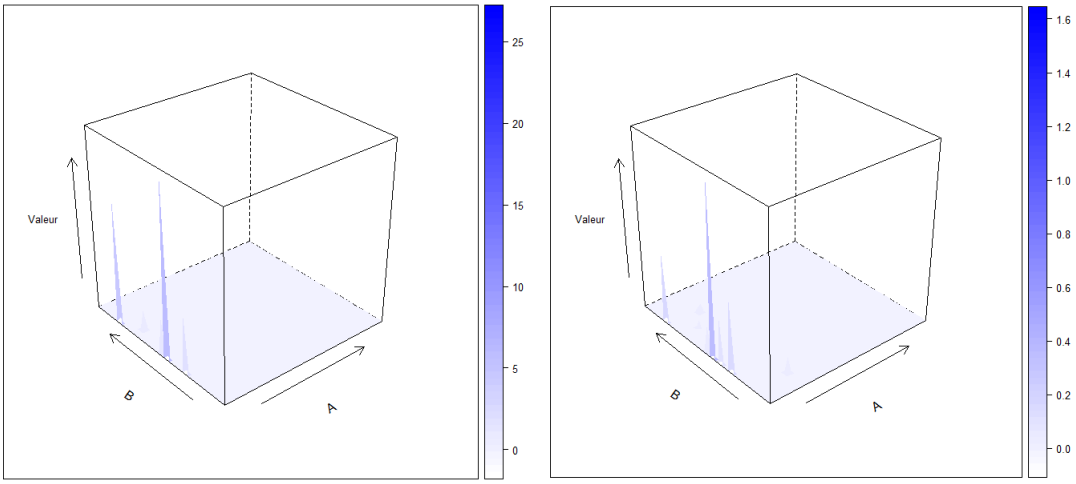
Deuxième logique de simulation

FIGURE 2.7 – Deuxième logique de localisation : Secteur A



(a) Valeur maximale de l'indice DO

(b) Valeur cumulée de l'indice DO



(c) Valeur maximale de l'indice M

(d) Valeur cumulée de l'indice M

Les résultats pour le secteur A sont proches des résultats de la première simulation. Dans ce cas, on peut dire que cela ne fait pas de différence que les localisations contrefactuelles soient réparties selon une répartition aléatoire ou de Poisson. Ce n'est pas le cas du secteur B pour lequel la sensibilité au nombre d'établissements est modifiée lorsque la répartition des établissements s'établit selon un processus de Poisson. Pour l'indice DO, on conclut plus souvent à de la concentration lorsqu'il y a plus d'établissements du secteur A que du secteur B. En revanche, pour l'indice M, il y a de la concentration quand il y a très peu d'établissements du secteur B, quel que soit le nombre d'établissements du secteur A.

Dans cette configuration particulière, l'indice M est toujours sensible au nombre d'établissements, alors que l'indice DO est plutôt sensible aux nombres relatifs des établissements des deux secteurs.

De la dispersion est également trouvée si le nombre d'établissements du secteur d'intérêt est faible.

Troisième logique de localisation : Secteur A

Comme pour la première configuration, les logiques de simulations sont similaires pour les secteurs A et B. Dans ce cas, pour les deux indices et d'autant plus pour l'indice DO, on trouve plus souvent de la concentration s'il y a peu d'établissements du secteur d'intérêt et beaucoup de localisations contrefactuelles possibles.

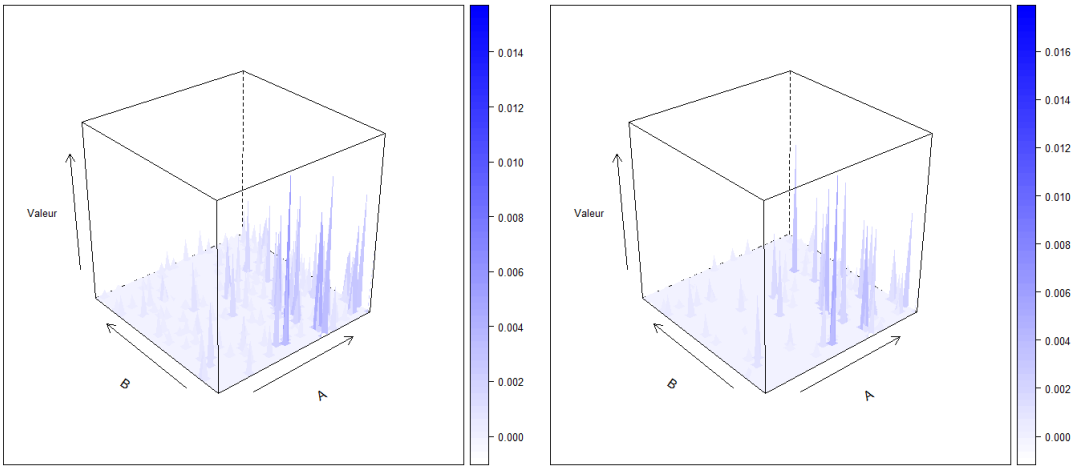
On peut également conclure à de la dispersion quand il y a peu d'établissements du secteur d'intérêt.

Il y a donc une sensibilité à la fois absolue et relative aux nombres d'établissements, ceux du secteur étudié et ceux des localisations contrefactuelles. Cette sensibilité n'est pas la même selon la logique de localisation des secteurs.

De plus, le Tableau 2.5 dénombre le nombre de fois où chaque indice indique de la concentration pour les 2500 combinaisons du nombre d'établissements des deux secteurs A et B, pour chaque simulation. Plutôt que de considérer la valeur de l'indice, on indique si celle-ci dépasse la limite supérieure de l'intervalle de confiance au moins une fois pour les distances considérées. Les résultats obtenus montrent que l'indice M conclut deux fois plus souvent à la concentration géographique des établissements que l'indice DO, quelle que soit la logique de localisation des établissements [13]. L'indice DO est donc plus robuste sur ce point que l'indice M.

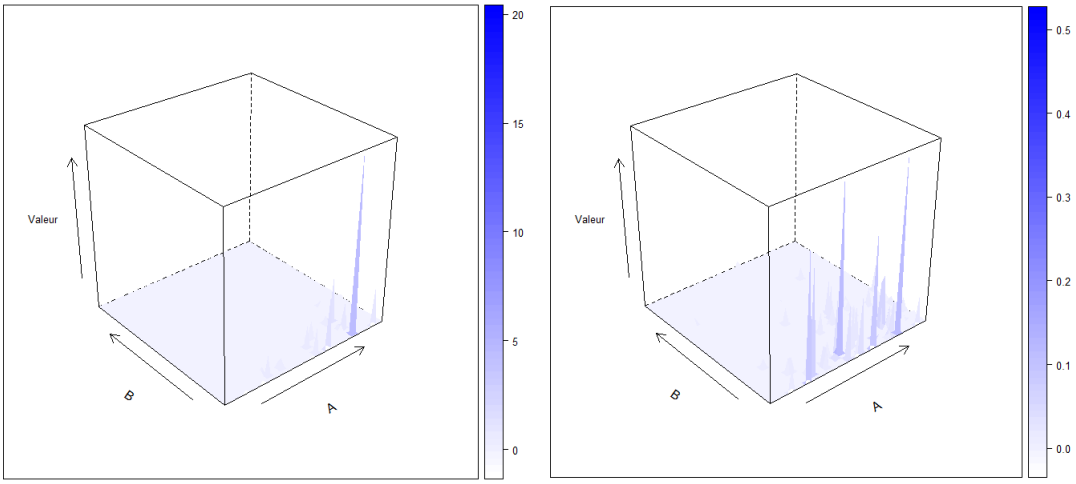
13. Pour les indices il n'y a pas de lien avec le nombre d'établissements de l'un ou l'autre des secteurs.

FIGURE 2.8 – Deuxième logique de localisation : Secteur B



(a) Valeur maximale de l'indice DO

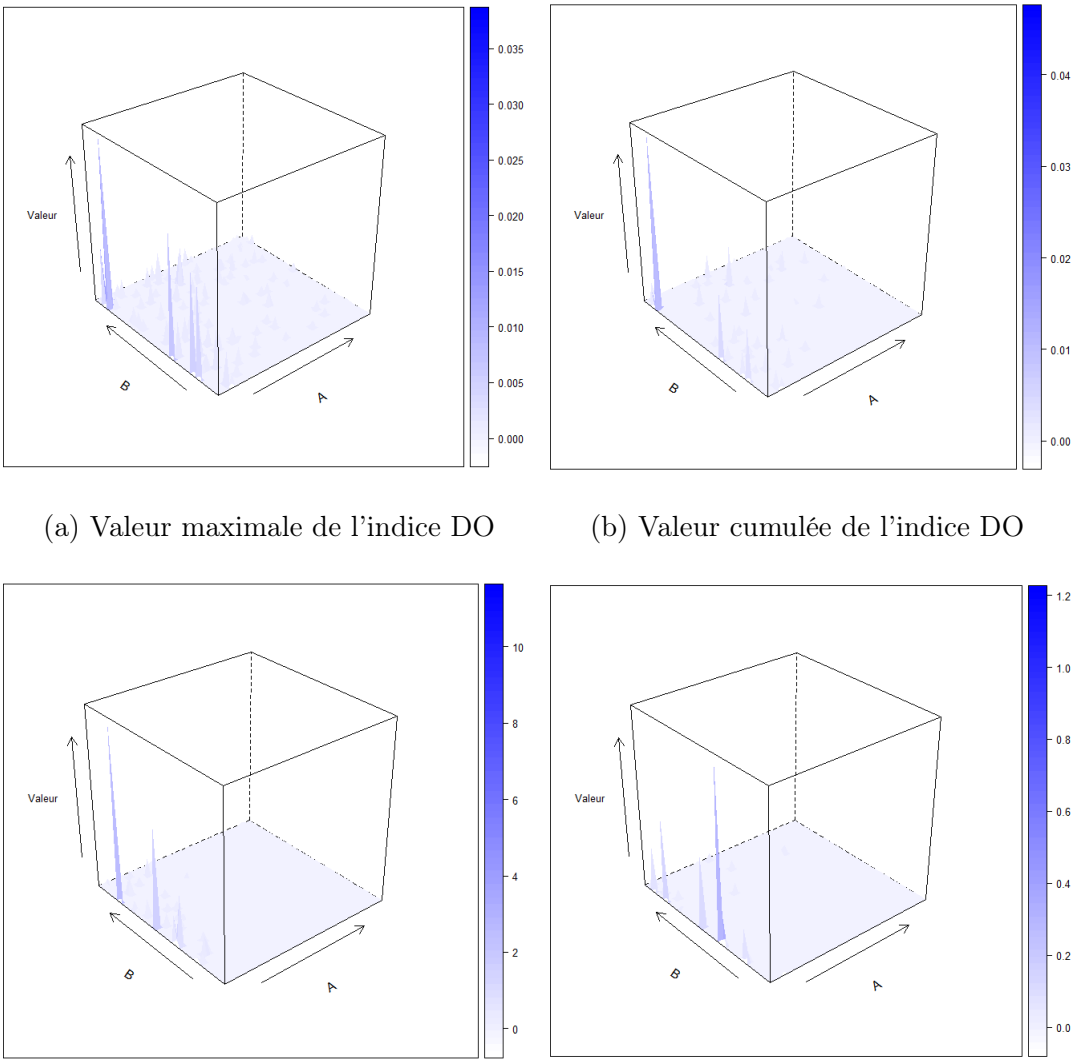
(b) Valeur cumulée de l'indice DO



(c) Valeur maximale de l'indice M

(d) Valeur cumulée de l'indice M

FIGURE 2.9 – Troisième logique de localisation : Secteur A



(a) Valeur maximale de l'indice DO

(b) Valeur cumulée de l'indice DO

(c) Valeur maximale de l'indice M

(d) Valeur cumulée de l'indice M

Tableau 2.5 – Dénombrement du nombre de concentration

	Simulation 1	Simulation 2	Simulation 3
MP_A	441	495	468
MP_B	432	458	446
DO_A	248	274	254
DO_B	254	277	251

2.4 Choix des localisations contrefactuelles et mise en évidence du MSUP

Le choix des localisations contre-factuelles utilisées pour construire les intervalles de confiance est crucial. Ce calcul repose sur des localisations possibles qui sont considérées identiques pour tous les établissements d'un même secteur à un niveau agrégé. De cette manière Marcon et Puech (2014) considère que les localisations possibles sont identiques pour les commerces d'habillement et les commerces de carburants car ces deux secteurs appartiennent au secteur "commerces de détail non alimentaires". Cela ne prend pas en compte que de nombreux facteurs très différents peuvent influencer le choix de localisation d'établissements d'un même secteur, cela étant d'autant plus vraisemblable que la nomenclature des activités est très agrégée. Par exemple, les établissements peuvent ne pas être sensibles aux même facteurs selon le nombre de salariés. De la même manière, les déterminants de localisation à l'échelle du secteur ne sont pas forcément identiques à ceux des sous-secteurs. De plus, les Tableaux 2.6 à 2.8 indiquent une grande diversité dans la construction de la nomenclature (nombre de sous-secteurs de NAF5 à NAF2 appartenant à un même secteur NAF4 à NAF1).

Pour vérifier si les logiques de localisations de sous-secteurs sont similaires, nous avons mesuré la coagglomération des établissements de tous les secteurs et sous-secteurs en utilisant la nomenclature des activités françaises (NAF) [14].

Nous pouvons ainsi savoir si la coagglomération est plus forte pour des sous-secteurs appartenant à un même secteur. Si oui, cela veut dire qu'on va prendre comme possibles des localisations qui sont proches de celles déjà choisies. Les indices continus concluront moins facilement à la concentration des établissements du secteur. Si non, cela veut dire qu'on va prendre comme localisations possibles des localisations qui sont éloignées de celles déjà choisies. Les indices continus concluront plus facilement à la concentration du secteur.

Sur la base de données 2014 de l'INSEE, nous considérons différentes combinaisons de décou-

14. Cette nomenclature des activités économiques productives a pour objectif de faciliter l'organisation de l'information économique et sociale. La NAF a la même structure que la nomenclature d'activités européenne NACE, elle-même dérivée de la nomenclature internationale CITI.

Tableau 2.6 – Récapitulatif de la nomenclature

	NAF5	NAF4	NAF3	NAF2	NAF1
Nombre de secteurs	732	615	272	88	21
Nombre de secteurs NAF4 avec un seul secteur NAF5	-	531	118	8	1
Nombre de secteurs NAF3 avec un seul secteur NAF4	-	-	138	8	1
Nombre de secteurs NAF2 avec un seul secteur NAF3	-	-	-	14	1
Nombre de secteurs NAF1 avec un seul secteur NAF2	-	-	-	-	5

Tableau 2.7 – Appartenance à la NAF des secteurs NAF5

	Nombre de secteurs sans autre secteur NAF5 dans la NAF du niveau supérieur	Médiane nombre de secteurs NAF5	Moyenne nombre de secteurs NAF5	Max nombre de secteurs NAF5	Écart type du nombre de secteurs NAF5
NAF4	531	2	2.39	6	0.84
NAF3	118	3	3.99	13	2.32
NAF2	8	6.5	9.05	59	9.73
NAF1	1	20	36.55	259	57.64

Tableau 2.8 – Appartenance à la NAF

	Nombre de secteurs sans autre secteur dans la NAF du niveau supérieur	Médiane nombre de secteurs	Moyenne nombre de secteurs	Max nombre de secteurs	Écart type du nombre de secteurs
NAF4	531	2	2.39	6	0.84
NAF3	138	3	3.56	9	1.96
NAF2	14	3	3.49	9	1.91
NAF1	5	3.5	5.19	24	5.23

pages géographiques et sectoriels. Puis, pour chaque combinaison nous calculons si la coagglomération est plus forte pour les sous-secteurs appartenant à un même secteur selon la nomenclature sectorielle.

Pour chaque secteur nous regardons la valeur de l'indice de coagglomération sectorielle d'Ellison et Glaeser selon l'équation :

$$\gamma_{ij} = \frac{\sum_{m=1}^{M}(s_{mi} - x_m)(s_{mj} - x_m)}{1 - \sum_{m=1}^{M} x_m^2} \tag{2.4}$$

où s_{mi} est la part de l'emploi i du secteur dans une zone m, et x_m la part de la zone m dans l'emploi de tout le territoire.

Pour chaque secteur nous pouvons vérifier si la coagglomération des secteurs appartenant à la même NAF est plus élevée que la coagglomération avec les autres secteurs. Si la coagglomération est élevée, alors l'utilisation de ces localisations comme contrefactuelles va conduire à conclure à

la dispersion des établissements ou à une répartition aléatoire. Si la coagglomération est faible, alors l'utilisation des localisations comme contrefactuelles conduira à conclure à la concentration géographique des établissements.

Tableau 2.9 – Rang de la coagglomération des secteurs NAF5

	Rang moyen	Rang minimum	Rang médian	Rang maximum	Écart type
Commune	10.65	1	2	221	24.91
Bassin de vie	104.15	1	89	378	95.59
Arrondissement	52.79	1	14	389	75.85
Zone d'emploi	80.93	1	41.5	385	93.66
Département	45.06	1	17	375	60.84
Région	83.05	1	62	349	78.46
Région 2016	85.24	1	64	370	77.72

Un premier travail descriptif sur données empiriques montre que la concentration des secteurs n'est pas la plus forte avec les établissements d'un même secteur. Même s'il y a toujours au moins un secteur pour lequel la concentration est la plus forte pour les établissements d'un même secteur, et ce quel que soit le découpage géographique considéré, les rangs médians et moyens montrent une forte variabilité au découpage géographique. Le rang moyen est de 10.65 pour le découpage Commune contre un rang moyen de 104.15 pour le découpage Bassin de vie. De même pour le rang médian qui est de 2 pour le découpage communal mais seulement de 89 pour le découpage Bassin de vie. Cette variabilité ne permet pas non plus d'établir un lien systématique entre le niveau d'agrégation et le niveau de coagglomération des établissements d'un même secteur.

Bien que ce résultat soit obtenu à partir d'une mesure discrétisant l'espace, cela montre également que le choix de la distance seuil peut conduire à des conclusions différentes. Nous verrons dans la suite l'effet de la modification de la distance seuil sur les mesures de concentration.

Tableau 2.10 – Rang de la coagglomération des secteurs NAF5, à la commune

	Nombre de secteurs à tester	Rang moyen	Rang moyen minimum	Rang moyen médian	Rang moyen maximum	Écart type du rang moyen
NAF4	201	135.14	1	88	672	139.38
NAF3	613	165.90	2	137.25	719	133.92
NAF2	719	199.17	2	185.04	721	125.63
NAF1	726	263.73	42.32	259.93	573.29	111.58

En étudiant la coagglomération des secteurs d'une même branche de la nomenclature des activités, on constate que la coagglomération moyenne diminue au fur et à mesure de l'agrégation sectorielle des données. Néanmoins, comme le nombre de secteurs testés augmente, cette baisse peut également signifier que l'agrégation sectorielle n'a pas de lien avec la localisation des acti-

vités. Cela valide l'hypothèse que la prise en compte de localisations contrefactuelles à partir des établissements appartenant à la même NAF aurait tendance à conclure plus souvent à la concentration du secteur. La création d'intervalles de confiance à partir de localisations contrefactuelles non spécifiques à chaque secteur constitue l'une des limite des indices de concentration continus car cette méthode est sensible au MSUP (Modifiable Sector-based Unit Problem) mentionné par Puech (2003).

Il nous apparait donc souhaitable de mener un travail préalable pour sélectionner plus finement les localisations contrefactuelles et l'échelle sectorielle adaptée. Les localisations contrefactuelles ne sont pas forcément les localisations choisies par les établissements de secteurs proches. En effet, il est facile d'imaginer que les localisations possibles évoluent au cours du temps et il apparait nécessaire de juger comme possibles des localisations partageant les mêmes caractéristiques au moment de l'implantation.

2.5 Conclusion

Les apports des méthodes continues sont théoriquement considérables pour les études impliquant économie et géographie, et plus particulièrement pour la localisation des activités. Ces méthodes permettent en effet d'en finir avec une vision discrète du territoire qui se résumait à un comptage d'observations dans chaque zone géographique. La prise en compte d'effet de voisinage est donc d'autant plus pertinente en privilégiant la distance entre établissements plutôt que l'appartenance (binaire ou contiguë) à des zones géographiques.

Néanmoins la mise en application de ces méthodes est confrontée à plusieurs difficultés. Ces dernières peuvent être liées aux données, particulièrement l'arbitrage à faire entre la précision géographique et le détails des observations. Les choix des différents paramètres des indices peuvent influencer les mesures de concentration et comme pour les indices discrets, les mesures continues sont sensibles au nombre d'établissements du secteur. De plus, il n'y a pas, à notre connaissance, de travaux s'intéressant au choix d'autres densités. Le seul travail relatif à la méthode de lissage de la fonction de densité est celui mené par Di Salvo *et al.* (2005). Les différentes fonctions se distinguent par la décroissance de la pondération au fur et à mesure que la distance entre les points s'accroit.

De nombreux travaux s'attachent à comparer la concentration des secteurs entre les différents pays. De ce point de vue, la catégorisation des établissements à partir d'une même classification sectorielle semble un prérequis. Ensuite, la comparaison de territoire alors que la proximité n'est pas appréhendée à partir du même choix de distance nécessite une clarification car même si les comparaisons qualitatives peuvent ne pas être affectées, les valeurs peuvent être sujettes à caution.

Enfin, la comparaison de secteurs n'ayant pas le même nombre d'établissements nous semble très délicate au regard des simulations considérées. En conclusion, une fois les difficultés d'accès aux données surmontées, il ne nous semble pas opportun que ces mesures soient systématiquement utilisées sans prendre de précautions préalable. Il est nécessaire de cibler des localisations d'intérêts (secteur, nombre d'employés par établissement, etc) qui ont une même sensibilité aux facteurs de localisations. Les distinctions prises en compte pour étudier la localisation des activités, services ou industries, et la comparaison des localisations à différentes années notamment, sont un premier pas vers une différenciation sectorielle des études. Ce n'est qu'après que ce travail de différenciation des secteurs ait été réalisé, que la sélection des localisations contrefactuelles et l'application des mesures continues peuvent se faire.

CHAPITRE 3

Modélisation empirique du choix de localisation : une étude sur données simulées à partir du modèle de Weber

Ce chapitre fait l'objet d'un travail co-écrit avec Salima Bouayad Agha (GAINS)

Résumé

Le choix de localisation d'un établissement est stratégique pour une entreprise. Identifier les principaux déterminants de ce choix, pour inciter les établissements à s'implanter est un enjeu majeur pour les décideurs publics et ce quelle que soit l'échelle de décision. Néanmoins, les études empiriques sur les choix de localisation des entreprises ne permettent pas forcément de formaliser le modèle sous-jacent. En effet, des facteurs multiples et parfois inobservables peuvent influencer la décision d'implantation. Dans ce chapitre nous désirons vérifier la capacité de certaines spécifications économétriques à restituer les déterminants d'un modèle de localisation inspiré du modèle de Weber. Pour cela notre approche consiste à simuler les choix de localisation selon un processus et à partir de déterminants que nous contrôlons, puis de vérifier si les estimations permettent de retrouver des résultats concordants avec les simulations de départ. Nous estimons à la fois des modèles de choix discrets (Logit, Probit et Logit conditionnel) ainsi que des modèles de comptage (Poisson, Binomial négatif). Cette approche nous permet également de vérifier dans quelle mesure les résultats des modèles les plus souvent utilisés pour identifier les facteurs de localisation sont sensibles au découpage géographique pris en compte et donc au MAUP (Modifiable Areal Unit Problem).

3.1 Introduction

Suite aux travaux de Krugman (1991a,b) et de la Nouvelle Économie Géographique (NEG) [1], le nombre d'études étudiant l'agglomération des activités a considérablement augmenté. Plus récemment, les travaux empiriques visant à identifier les déterminants de la localisation des entreprises ont également connu un regain d'intérêt. Ces études bénéficient à la fois du développement concomitant de travaux théoriques (Fujita *et al.*, 1999; Fujita et Thisse, 2002; Krugman, 1991a), de méthodes économétriques adaptées (McFadden, 1974), ainsi que d'une disponibilité croissante de données géolocalisées. Cette recherche des déterminants de localisation les plus significatifs est aussi à mettre en lien avec la mise en œuvre de programmes et de politiques publiques visant à attirer et à promouvoir la création et l'implantation de nouvelles activités sur les territoires. Dans une conjoncture économique difficile, cibler les projets les plus bénéfiques devient un enjeu crucial pour les territoires. D'autant plus en période de crise pendant laquelle les fonds alloués à l'implantation d'entreprises sont plus faibles. Mieux connaître les territoires et identifier les mécanismes à l'œuvre au moment de la décision du choix de localisation permet de disposer d'informations cruciales tant pour les entreprises que pour les décideurs publics.

Pour une entreprise, disposer d'informations sur les caractéristiques des territoires facilite la prise de décisions en matière d'implantation, de délocalisation ou de consolidation de son activité (agrandissement d'un site par exemple), sur les zones les plus favorables à celle-ci. Les pouvoirs publics locaux peuvent quant à eux cibler leur politique en direction des entreprises les plus sensibles aux facteurs présents sur leur territoire. L'identification des déterminants les plus significatifs des choix de localisation doit ainsi permettre de mettre en place des politiques publiques plus efficaces. Des politiques généralistes d'incitation à la création d'entreprises sur un territoire donné, sans qu'il n'y ait de ciblage préalable, sont sources d'incertitudes sur le nombre et le type d'entreprises qui souhaitent s'implanter. Le risque d'effet d'aubaine est coûteux à court terme en raison d'aides accordées à des entreprises qui ne sont pas capitales pour le territoire, mais également à long terme car ces entreprises ne vont pas créer les richesses nécessaires à leur développement. Des pertes peuvent aussi survenir si des entreprises profitent d'aides territoriales alors qu'elles auraient de toutes les manières pris la décision de s'y implanter même sans aides.

Comme le choix de sa localisation est une décision stratégique pour une nouvelle entreprise, il est important d'évaluer empiriquement les facteurs qui déterminent cette décision et de le faire en retenant la spécification la plus appropriée. En se basant sur un modèle qui simule la localisation à

1. Voir les discussions de Duranton (1997),Coissard (2007) et Crozet et Lafourcade (2010) pour plus de détails sur la NEG.

partir de la distance à des facteurs choisis arbitrairement, nous allons identifier quelle spécification serait la plus pertinente pour retrouver les facteurs déterminants . Pour cela, on peut envisager une approche en termes de choix discret. Celle-ci distingue les facteurs se rapportant à l'agent qui prend la décision, de ceux liés à l'ensemble des possibilités à partir duquel le choix est opéré (c-à-d. le territoire, la zone spatiale). On peut aussi aborder les choix de localisation du point de vue de l'agent qui fait le choix, ou sous l'angle du territoire choisi. Les spécifications économétriques sont différentes selon le point de vue envisagé. Lorsque l'entreprise constitue l'unité d'analyse et que la préoccupation principale est de savoir comment ses caractéristiques (taille, secteur, etc) ou celles du territoire choisi (population, infrastructures, etc) affectent les décisions de localisation, ce sont les modèles à choix discrets (MCD ci-après) qui sont utilisés. En revanche, si l'unité d'analyse est géographique (municipalité, comté, province, région, etc) et que les facteurs affectant les décisions de localisation font référence au territoire, on considère des modèles pour données de comptage (MDC ci-après).

Dans les MCD, on suppose qu'une entreprise prend une décision rationnelle basée sur la maximisation de son profit (ou la minimisation de ses coûts) afin que les avantages accumulés dépassent les investissements initiaux en capital ainsi que les dépenses organisationnelles qui en découlent. Plus précisément, ces modèles permettent d'évaluer l'influence des caractéristiques du décideur sur le choix d'implantation (telles que la taille de l'entreprise, le secteur d'activité, etc), en prenant en compte l'ensemble des alternatives géographiques disponibles (McFadden, 1974).

Les MDC étudient l'impact des attributs de localisation des firmes, en prenant en compte le nombre d'entreprises dans chaque unité territoriale. Ces modèles traitent la question du choix de localisation en recherchant les caractéristiques d'une zone qui affectent le nombre d'entreprises établies à un endroit précis et à une période donnée (Henderson et Becker, 2000) [2].

Les déterminants des choix d'implantation des entreprises sont très souvent liés aux caractéristiques d'une zone (accès préférentiel aux réseaux de communications, forte agglomération d'activité, taux de chômage local) et peuvent être spécifiques à une industrie (présence d'un cluster du même secteur, territoire clé pour l'industrie, main d'œuvre adaptée, fiscalité avantageuse). Comme ces déterminants sont spatialisés, il convient de choisir le niveau géographique le plus adéquat pour chacun d'entre eux. Afin d'étudier les facteurs qui motivent l'implantation d'une entreprise à un endroit plutôt qu'à un autre, l'espace est nécessairement discrétisé et les données mobilisées pour ces études peuvent parfois être issues de découpages géographiques différents. Le concept de MAUP (Modifiable Areal Unit Problem), introduit par Openshaw et Taylor (1979, 1981) et Openshaw (1984), désigne l'influence du découpage spatial sur les résultats. Il s'agit de la combinaison de deux pro-

2. Voir Arauzo-Carod *et al.* (2010) pour un détail des avantages et inconvénients de ces méthodes.

blèmes distincts mais proches : un problème d'agrégation (lié à un changement dans la diversité de l'information engendré par les différents schémas possibles d'agrégation à une même échelle) et un problème d'échelle (lié à une variation de l'information engendrée lorsqu'un jeu d'unités spatiales est agrégé afin de former des unités moins nombreuses et plus grandes).

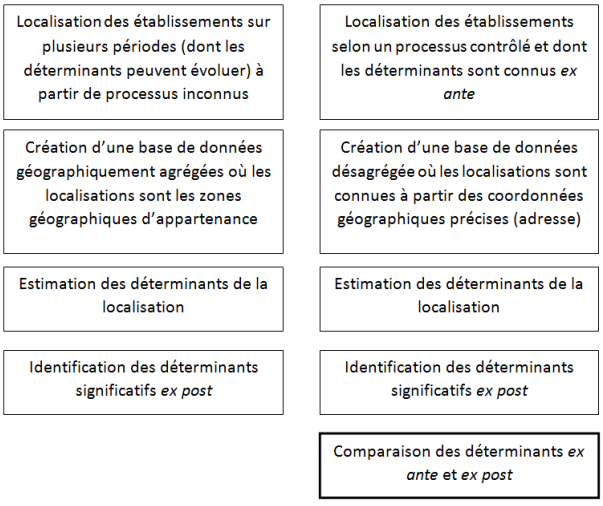
Comme cela est montré par Arauzo-Carod (2008), le niveau de découpage d'une variable (expliquée ou explicative) est crucial car il peut dépendre d'une politique publique à un certain niveau géographique. Il est donc important de relier la variable explicative à l'échelle géographique à laquelle les décideurs locaux sont compétents. Cependant, cela n'est pas possible dans tous les cas puisque les données ne sont pas toujours accessibles ou existantes. Notons également qu'il est possible de mélanger des variables issues de différents découpages géographiques comme le fait Woodward (1992).

Dans ce travail, nous étudions la capacité des modèles de choix de localisation à restituer les déterminants d'implantations préalablement contrôlés. Pour cela, nous simulons des localisations à partir d'un modèle de choix pour lequel les déterminants sont prédéterminés mais ne dépendent pas du découpage géographique. Nous étudions ensuite dans quelle mesure ces facteurs sont jugés comme significatifs en considérant plusieurs spécifications économétriques. Plus précisément, nous considérons un modèle à la Weber (1909) pour lequel le choix de localisation résulte de la minimisation du coût total de transport entre l'entreprise, les localisations d'input et le marché final. Ce modèle repose sur la distance euclidienne et ne dépend donc pas du découpage géographique adopté, il n'est pas sensible au MAUP. Ce type d'approche repose sur une vision continue de l'espace, comme c'est notamment le cas de la littérature étudiant la concentration géographique des établissements à partir d'indices continus. Dans cette littérature, les indices prenant explicitement en compte la distance entre les établissements (Duranton et Overman, 2005; Marcon et Puech, 2003) améliorent les informations obtenues à partir d'indices discrétisant l'espace en zones (Ellison et Glaeser, 1994; Maurel et Sédillot, 1999). L'hypothèse de continuité de l'espace modifie donc la manière dont on appréhende le choix de localisation car une entreprise se localise à un endroit plutôt que dans une zone qui lui permet de minimiser ses coûts de transports. En revanche, cela suppose de disposer de données désagrégées qui ne sont pas toujours disponibles. C'est sans doute pour cette raison que la plupart des études empiriques sur les choix de localisation reposent sur des données agrégées. On notera cependant le travail d'Arbia (2001) qui tente d'expliquer la localisation et l'agglomération des activités à San Marin en considérant l'espace comme continu. D'autre part, on peut considérer la distance comme l'une des variables explicatives de la localisation (des établissements industriels (Smith et Florida, 1994) ; des établissements du secteur automobile Klier

et McMillen (2008, 2015)). Ces variables de distance sont le plus souvent calculées à partir de points considérés pour l'ensemble d'un territoire (distance entre les capitales de pays comme équivalent à la distance entre deux pays), ou à partir d'une distance moyenne. De plus, elles ne sont pas testées seules mais en complément de variables plus classiques (agglomération des entreprises, densité de population, qualité de la main d'œuvre, etc).

Dans ce travail nous allons chercher à évaluer la qualité des résultats de différentes spécifications économétriques à partir de localisations déterminées selon un processus contrôlé. Nous pouvons ainsi vérifier si les facteurs préalablement retenus comme déterminants du choix de localisation sont ceux ayant un impact significatif lors de l'estimation. En procédant à une analyse exhaustive des écarts qui séparent les résultats estimés des paramètres utilisés pour simuler les localisations, nous identifions la sensibilité de chaque spécification à chaque facteur de localisation. La Figure 3.1 synthétise ce qui différencie notre approche d'une démarche traditionnelle. L'objectif du chapitre n'est pas de déterminer quels sont les facteurs du choix de localisation, mais si les estimateurs des paramètres d'un modèle de choix de localisation sont conformes à ceux qui ont été simulés à partir dun modèle théorique.

FIGURE 3.1 – Démarche traditionnelle (à gauche), démarche du chapitre (à droite)



Dans une deuxième partie, nous présentons le modèle à partir duquel nous simulons des localisations d'établissements et le protocole retenu pour générer les données. Nous présentons ensuite les variables et spécifications utilisées pour estimer les effets des déterminants de la localisation. La quatrième section analyse les résultats et notamment leur sensibilité aux différents découpages géographiques, avant de conclure quant aux spécifications à privilégier.

3.2 Simulations des choix de localisation

Dans cette section, nous présentons le modèle initial de Weber à partir duquel nous allons simuler les choix de localisation. Il s'agit d'abord de définir les déterminants des implantations, puis de calculer le coût associé à chacune d'entre elles et dans une dernière étape, de localiser les établissements du secteur à étudier.

3.2.1 Modèle de Weber

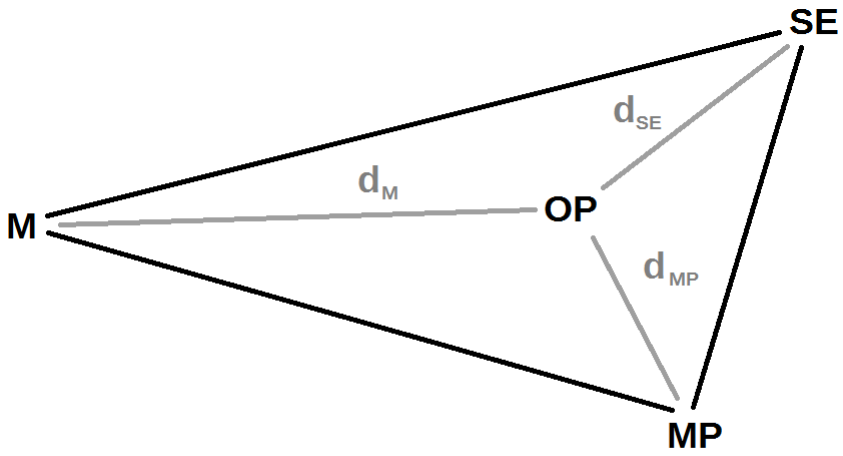
Le modèle de Weber (1929) s'intéresse à ce que serait la localisation optimale (OP) d'une entreprise dans un espace parfaitement homogène où se trouvent trois points correspondant à la source d'énergie (SE), au fournisseur de matières premières (MP) et au marché (M). Sous certaines conditions, le modèle démontre que le coût de transport détermine la localisation idéale de l'établissement. Plus précisément, le coût total de transport dépend de la quantité et du coût de transport unitaire des inputs et de l'output, de la localisation des fournisseurs d'inputs et du marché final. Une cartographie du territoire permettant de rendre compte du coût associé à chaque localisation peut ainsi être réalisée pour distinguer les zones minimisant le coût total de transport. La localisation optimale correspond géométriquement au barycentre du triangle formé par la localisation des inputs et du marché (que l'on peut pondérer si le coût de transport n'est pas le même pour les inputs et outputs). Plus un facteur est important pour l'entreprise, plus l'entreprise se rapprochera géographiquement de ce facteur.

De manière générale, le problème consiste à maximiser le profit de l'entreprise Π. Cela revient également à minimiser le coût total de transport, noté CT, une fois les quantités d'inputs et d'outputs ainsi que la localisation des deux fournisseurs et du marché final déterminées. Le travail de Weber se focalise sur la minimisation du coût total de transport, c'est à dire la localisation qui minimise ce coût compte tenu de la localisation des deux inputs (SE et MP) et du marché sur lequel vendre l'output (M). Toutes les localisations possibles sont caractérisées par la distance avec la localisation choisie (soit $d_{SE}, d_{MP} et d_M$) ainsi que par la quantité d'inputs utilisés (q_{SE} et q_{MP}) et de leur prix de transport respectif (Pt_{SE}, Pt_{MP} et Pt_M), ainsi que par la quantité d'outputs produits (q_M). À ceci s'ajoute le prix d'achat des inputs (P_{SE} et P_{MP}) et le prix de vente de l'output (P_M).

Ainsi, la localisation optimale doit permettre de minimiser le coût total : $CT = P_{SE} \times q_{SE} \times Pt_{SE} \times dist_{SE} + P_{MP} \times q_{MP} \times Pt_{MP} \times dist_{MP} + P_M \times q_M \times Pt_M \times dist_M$

La substituabilité des facteurs de production peut également impacter le choix d'implantation (Moses, 1958). Si les inputs sont substituables, l'entreprise peut augmenter la quantité d'un des

FIGURE 3.2 – Représentation du triangle de Weber



deux inputs pour compenser par exemple une plus grande distance à l'autre input, ou un coût de transport unitaire plus élevé.

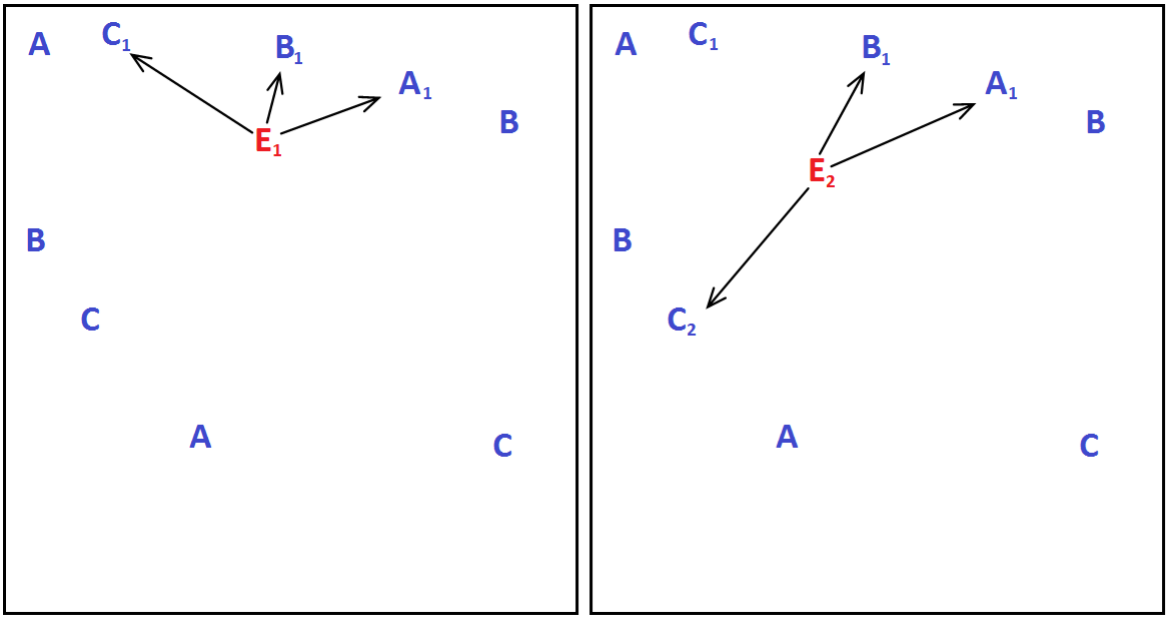
Ce modèle peut être étendu en considérant des zones pour lesquelles la présence d'une main d'œuvre qualifiée permet de réduire le coût du processus de production. De même, des zones permettant de bénéficier d'économies d'agglomération peuvent expliquer l'écart entre la localisation minimisant le coût total de transport et la localisation choisie.

Dans le modèle initial, l'entreprise qui cherche à s'implanter connaît déjà la localisation des fournisseurs et du marché final qui sont prédéterminées ou qui résultent d'un choix précédent. Dans le modèle sur lequel reposent les simulations des coordonnées des établissements, l'entreprise qui cherche à s'implanter connaît la localisation et le type de toutes les autres entreprises. Néanmoins, les entreprises auxquelles elle fera appel dépendent de la localisation qu'elle choisira. Selon l'endroit considéré sur le territoire, l'entreprise ne fera pas appel au même fournisseur. La Figure 3.3 indique les fournisseurs les plus proches pour deux localisations potentielles d'un établissement du secteur E (E_1 et E_2). Selon la localisation envisagée, les entreprises des secteurs A et B sont identiques (A_1 et B_1) car les plus proches. En revanche, l'entreprise du secteur C qui sera choisie ne sera pas la même suivant les deux localisations envisagées, C_1 pour une localisation en E_1 et C_2 pour une localisation en E_2, car la distance $[E_2C_1]$ est supérieure à $[E_2C_2]$.

Le modèle de Weber est un modèle de choix de localisation simplifié ; tous les paramètres sont connus au moment du choix de localisation et le modèle repose sur l'hypothèse que la localisation qui minimise le coût de transport est également celle qui maximise le profit. Or, cette hypothèse n'est valable que si la fonction de production de la firme induit une complémentarité des facteurs (fonction de la forme Walras-Leontief)[3]. De plus, selon Moses (1958) la localisation optimale ne

3. Peeters et Perreur (1996) listent les conditions nécessaires pour que cette hypothèse soit vérifiée et approfondissent la discussion sur la localisation wébérienne des activités.

FIGURE 3.3 – Exemple de choix des fournisseurs pour deux localisations proches



doit pas dépendre du niveau d'outputs produit : la localisation optimale pour la production d'un seul output est identique à celle obtenue à partir d'une infinité d'outputs.

Ce modèle de localisation ne prend pas en compte qu'une entreprise puisse anticiper l'accès à de nouveaux marchés ou solliciter d'autres inputs, cela découle d'une vision statique. Il n'est donc pertinent que pour une localisation à un instant donné. Au fur et à mesure du temps, une localisation optimale peut ne plus l'être (du fait de l'évolution du coût de transport ou du réseau routier), mais l'entreprise peut néanmoins rester au même endroit si les coûts engendrés par un déménagement sont plus importants que les pertes liées à la localisation non optimale. Cette éventualité remet également en cause les études des déterminants de choix de localisation si l'implantation de l'établissement ne correspond pas à la période de collecte des données des facteurs supposés déterminants du choix de localisation.

Malgré ces limites, le modèle de Weber reste adapté pour ce travail car il se base uniquement sur la distance entre localisations. Des améliorations peuvent lui être apportées mais nous nous limitons ici à une version alternative et simplifiée qui est suffisante pour ce travail.

3.2.2 Processus de simulations

Nous considérons jusqu'à trois inputs A, B et C nécessaires à la production d'un output pour les firmes d'un secteur E. Chaque input n'est produit que par un seul secteur et chaque secteur ne produit qu'un seul input. Que l'on soit dans le cas de deux inputs et un marché ou que l'on considère trois inputs, la recherche de la localisation optimale obéit aux même règles puisque nous

supposons que la quantité d'inputs disponible est donnée et ne dépend pas de la localisation. Les simulations qui nous permettent d'évaluer la pertinence des estimations obtenues à partir de différentes spécifications ne reposent pas sur la formalisation des fonctions de production. Pour chaque spécification, nous testons si la distance la plus courte aux inputs de chaque secteur favorise la localisation des établissements du secteur E. Afin de simplifier l'étude, nous supposons que les prix à l'achat et le coût de transport des trois inputs sont égaux à 1.

Pour simuler les localisations d'établissements du secteur d'intérêt, nous procédons en trois étapes :

1. Répartition géographique des inputs
2. Calcul de la probabilité d'implantation en chaque point du territoire
3. Tirage des localisations des établissements

Répartition géographique des inputs

Dans un premier temps, nous répartissons les trois types d'inputs sur un territoire supposé être un carré de côté égal à 1. Les établissements d'un secteur peuvent être répartis de manière aléatoire ou de manière agglomérée. Dans le premier cas, les coordonnées des établissements sont obtenues à partir d'un tirage aléatoire. Dans le second cas, les coordonnées des établissements sont obtenues à partir d'un processus de Matérn qui permet d'obtenir une répartition agglomérée des unités de production en un ou plusieurs pôles.

Calcul de la probabilité d'implantation en chaque point du territoire

L'équation de coût total pour un établissement du secteur E situé au point j s'écrit :

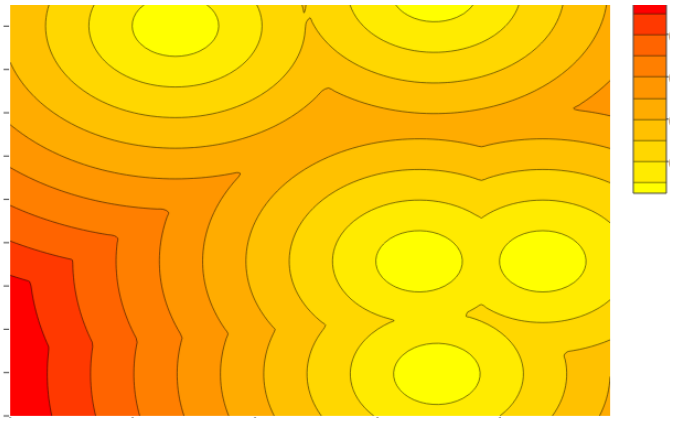
$CT_{E,j} = \beta_{E,A} \times dist_{j,A}^{\alpha} + \beta_{E,B} \times dist_{j,B}^{\alpha} + \beta_{E,C} \times dist_{j,C}^{\alpha}$

Soit : $CT_{E,j} = \sum_{f=A}^{C} \beta_{E,f} * dist_{j,f}^{\alpha}$

$\beta_{E,f}$ est le coefficient mesurant l'impact du facteur f sur le coût de production des établissements du secteur E. $dist_{j,F}$ est la distance minimale qui sépare le point j du facteur F. Le paramètre α permet d'introduire une variation non linéaire du coût total selon la distance. Selon cette spécification, le coût marginal de la distance est strictement croissant, $\alpha > 1$.

À partir de cette équation, il est possible de déterminer en tout point du territoire, le coût total qui résulterait d'une localisation potentielle en un point donné j. La Figure 3.4 représente un exemple de coût total de transport associé à chaque localisation possible du territoire. On constate plusieurs zones au coût de transport faible. Les établissements du secteur E auront plus de chances de se localiser dans ces zones. Dans cet exemple, les établissements du secteur d'intérêt auront tendance à se répartir sur une grande surface du territoire car les zones sont relativement éloignées.

FIGURE 3.4 – Représentation du coût de transport pour un établissement du secteur E en chaque point du territoire



Nous déterminons *a priori*, pour chacun des points du territoire, le nombre d'établissements du secteur E ($ETAB_E$). Puis, nous calculons en chaque point j du territoire, la probabilité qu'un établissement du secteur E s'y localise. Cette probabilité est inversement proportionnelle au coût total au point j. Plus le coût total est élevé au point j, plus la probabilité qu'un établissement du secteur E s'y localise est faible. Nous normalisons le coût de transport en chaque point entre 0 et 1 de telle sorte qu'il existe au moins un point où le coût total est de 0 et un point où le coût total est de 1. Nous considérons ensuite que le profit en chaque point est égal à 1 diminué du coût de transport en ce point. Le profit est également compris entre 0 et 1 en chaque point.

À partir de la fonction de répartition empirique de la distribution des profits, nous calculons la probabilité de localisation en chaque point. Du fait de notre processus de localisation, nous prenons en compte que la probabilité de localisation en un point dépend du profit réalisable en ce point et des profits possibles sur l'ensemble du territoire [4].

Tirage des localisations des établissements

En chaque point j, nous connaissons la probabilité d'implantation associée et nous utilisons la loi de Bernouilli pour implanter (ou non) un établissement du secteur E. En procédant ainsi, on obtient la localisation des entreprises du secteur E sur tout le territoire. Comme nous utilisons des probabilités de localisations, le nombre d'établissements du secteur E peut être différent du nombre préalablement fixé. Dans ce cas, nous répétons le tirage jusqu'à obtenir le nombre d'établissements désiré. La localisation au point j dépend du profit réalisable en ce point et des profits réalisables dans toutes les autres localisations possibles.

Une fois ces localisations simulées, il est possible d'estimer les modèles empiriques de choix de

4. Une autre méthode qui consisterait à choisir un certain nombre de localisations ayant le plus faible coût de transport peut conduire à une concentration des firmes autour d'un même point.

localisation et de comparer les valeurs des paramètres issus de ces estimations avec celles retenues préalablement pour simuler les implantations des établissements du secteur E.

Nous faisons varier les différents paramètres du modèle de localisation pour vérifier l'impact sur les spécifications testées :

— les logiques de localisations de chaque secteur (aléatoire ou agglomérée),

— les paramètres associés à chaque secteur (0, 1/3 , 2/3 et 1). Nous imposons que la somme des paramètres associés à chaque secteur est égale à 1. Si le paramètre associé à un secteur est nul, cela signifie que l'input n'est pas nécessaire aux entreprises du secteur E, tandis que si le paramètre est égal à 1 ce secteur est le seul déterminant de localisation. Les paramètres peuvent ainsi être interprétés comme des poids associés à chaque secteur d'inputs,

3.3 Données

Par construction des simulations, nous savons s'il y a un établissement localisé en chaque point du territoire. Nous pouvons ainsi créer plusieurs bases de données afin de pouvoir tester plusieurs modèles d'estimation (Tableau 3.1). En effet, à partir des simulations de localisations, les observations peuvent être des localisations choisies et non choisies (la variable dépendante est binaire) ou les observations sont agrégées à l'échelle des territoires (la variable dépendante est le nombre d'établissements du secteur étudié dans la zone).

Tableau 3.1 – Détails des modèles estimés et des variables explicatives selon la base de données

Localisations	Logit, Probit, Logit conditionnel, Binomial négatif, Poisson	Distance minimale à chaque type d'input
Agrégation	Poisson	Nombre de chaque type d'input dans la même zone géographique

Les spécifications retenues pour identifier les déterminants de la localisation sont les suivantes : logit, probit, logit conditionnel, binomial négatif, Poisson. Les trois premiers cités sont des MCD et les deux suivants des MDC. Pour les MCD, les observations correspondantes dans les données sont les choix de localisations possibles et choisies des entreprises, alors que les observations correspondent aux territoires pour les MDC.

Pour les modèles logit et probit, nous prenons comme observations l'ensemble des localisations possibles (choisies et non choisies) sur le territoire. La variable expliquée est égale à 1 si un établissement est implanté à cette localisation, 0 sinon. Les variables explicatives sont les distances les plus proches aux trois inputs. Pour le logit la fonction de répartition correspond à une loi logistique, alors que pour le probit la fonction de répartition suit une loi normale centrée réduite.

La loi logistique tend à attribuer aux événement extrêmes une probabilité plus forte que la distribution normale (Amemiya, 1981). Pour le logit conditionnel, nous déterminons neuf alternatives de localisations tirées au hasard parmi les localisations non choisies. Cette méthode est proche de celle de Ben-Akiva *et al.* (1985), utilisée également par Klier et McMillen (2008, 2015), ces derniers n'utilisent que cinq localisations alternatives pour les travaux cités. Il est toutefois possible que ces localisations soient proches donc que ces localisations alternatives soient corrélées. Notre base est donc composée de 1 000 localisations, 100 choisies et 900 non choisies.

Pour les MDC, nous utilisons les modèles de Poisson et binomial négatif. La deuxième alternative est préférable à la première en cas de surdispersion (si la variance est supérieure à l'espérance). Un degré de liberté supplémentaire est introduit par l'intermédiaire d'un facteur d'hétérogénéité (alors que ce facteur est nul pour le modèle de Poisson).

Nous comparons les valeurs des coefficients issues des estimations avec celles des vrais paramètres qui ont servi pour simuler les coordonnées. Selon les cas, on peut se retrouver dans une situation pour laquelle un facteur considéré comme déterminant dans le choix de localisation le soit également dans le modèle estimé (significatif et de bon signe) ou ne le soit pas (non significatif ou significatif mais de mauvais signe). On peut également se retrouver dans la situation où un facteur considéré comme non déterminant dans le modèle simulé (qui devrait être non significatif dans le modèle estimé) soit significatif selon le modèle estimé. Nous utiliserons un seuil de significativité de 5%. Une estimation est conforme si les déterminants sont significatifs et de bon signe, et les non déterminants sont non significatifs. Dès lors qu'au moins un autre cas est trouvé, l'estimation ne restitue pas au moins un paramètre du modèle de base et est considérée comme erronée. Compte tenu du modèle utilisé pour implanter les établissements, nous formulons trois propositions qui nous servirons de grille d'analyse des résultats.

Proposition 1 : Plus la localisation des inputs est aléatoire, plus il est difficile de retrouver les paramètres du modèle simulé.

Si les établissements d'un même secteur sont agrégés, les localisations favorables seront rapprochées géographiquement. Inversement, si les établissements d'un même input sont disséminés sur le territoire, les localisations favorables le seront également.

Proposition 2 : Plus les paramètres associés aux inputs sont élevés dans le modèle simulé, plus les coefficients estimés sont élevés.

Plus le paramètre associé est élevé, plus le coût de transport augmente donc plus le profit diminue. L'augmentation de la valeur du paramètre doit donc conduire à améliorer la significativité mesurée de l'input.

Proposition 3 : Plus les données sont agrégées plus il est difficile de retrouver le modèle de localisation.

Moins il y a de zones, plus les zones partagerons des caractéristiques communes, et plus il sera difficile de distinguer un territoire d'un autre. Il sera donc moins aisé d'identifier les déterminants de la localisation.

3.4 Résultats

Dans cette section, nous présentons dans un premier temps les résultats globaux. Puis, nous détaillons l'analyse des résultats pour chaque paramètre du modèle de localisation.

Nous procédons à 80 simulations de localisations. Pour évaluer la sensibilité au découpage géographique utilisé, nous procédons pour chacune de ces 80 simulations à un découpage en 4, 7, 10, 13, 16, 19 et 22 régions. Ces découpages sont obtenus à partir de diagramme de Voronoï. Pour ne pas dépendre d'un seul diagramme pour un découpage géographique nous produisons vingt découpages différents pour un nombre de régions donné. Le Tableau 3.2 présente le nombre d'estimations conformes au modèle de localisation pour les deux types de spécifications (MCD et MDC). L'estimation peut présenter au maximum trois erreurs relativement aux paramètres du modèle simulé, une fois pour chaque secteur d'input. Les spécifications sont assez proches même si le modèle simulé est un peu moins bien retrouvé pour le logit conditionnel. Les paramètres issus du modèle simulé est tout de même retrouvé dans environ 59% des cas (plus de 47 cas sur 80). Néanmoins, ce tableau n'indique pas quelles sont les erreurs (des déterminants non significatifs ou des non déterminants significatifs par exemple).

Tableau 3.2 – Nombre de cas conformes par modèle

	Sans erreur	Une erreur	Deux erreurs	Trois erreurs
MCD				
Logit	47	32	1	0
Probit	47	32	1	0
Logit conditionnel	43	36	1	0
MDC				
Poisson	47	32	1	0
Binomial négatif	47	32	1	0

Le Tableaux 3.3 détaille les résultats pour chaque déterminant et non déterminant pour les différentes spécifications. Il y a 144 déterminants et 96 non déterminants. Les résultats indiquent que les erreurs proviennent quasi-exclusivement de la non significativité d'un déterminant plus que de la significativité d'un non déterminant (plus de 30 cas contre 2). Les résultats sont très similaires

selon les différentes spécifications à l'exception du logit conditionnel qui trouve moins souvent les déterminants comme significatifs.

Tableau 3.3 – Détails des cas par modèle

	Déterminant			Non déterminant		
	Significatif	Non significatif	Erreur de signe	Significatif et positif	Non significatif	Significatif et négatif
MCD						
Logit	111	33	0	1	94	1
Probit	111	33	0	1	94	1
Logit conditionnel	107	36	1	1	92	3
MDC						
Poisson	111	33	0	1	94	1
Binomial négatif	111	33	0	1	94	1

3.4.1 Impact des paramètres de localisation

Les premiers résultats indiquent que certaines spécifications pouvaient réagir de la même façon suite à la modification d'un paramètre du modèle de localisation. Nous allons donc identifier quels sont les paramètres (logique de localisation, paramètres associés et nombre de zones géographiques) les plus impactants sur chaque spécification testée.

Logique de localisation

Nous commençons par vérifier si la logique de localisation des secteurs d'inputs peut influencer la capacité des modèles estimés à restituer les déterminants du modèle de localisation. L'objectif est de savoir si les estimations sont plus souvent conformes quand les établissements sont agglomérés. À l'inverse, si les localisations sont aléatoires et homogènes, on s'attend plutôt à ce qu'il soit difficile de bien retrouver les déterminants et non déterminants du modèle de localisation. Le Tableau **??** indique la part des simulations bien retrouvées en fonction du nombre de secteurs d'inputs concentrés. Les résultats sont relativement similaires d'un modèle à l'autre et n'indiquent pas une relation strictement croissante ou décroissante, même si le modèle de localisation est moins bien trouvé lorsque tous les secteurs sont concentrés géographiquement.

Ce résultat étant surprenant, nous avons procédé à une deuxième simulation de localisations. Lors de cette deuxième simulation, les résultats sont similaires lorsqu'il n'y a pas tous les secteurs concentrés. Quand les trois secteurs sont concentrés (10 cas), le modèle de localisation est bien retrouvé dans 70% des cas. Cela s'explique par la localisation des agrégats d'établissements. Quand les agrégats sont proches, il y a une coagglomération des établissements des secteurs et il est plus difficile de retrouver le bon modèle car les variables explicatives (les distances) sont similaires, ce qui

Tableau 3.4 – Part des estimations conformes selon le nombre de secteurs déterminants (en %)

	1	2	3
Nombre de cas	24	48	8
MCD			
Logit	95.83	50.00	0.00
Probit	95.83	50.00	0.00
Logit conditionnel	87.50	45.83	0.00
MDC			
Poisson	95.83	50.00	0.00
Binomial négatif	95.83	50.00	0.00

est le cas ici, un non déterminant peu ainsi être trouvé significatif et favorable à la localisation alors qu'il n'a aucun rôle dans la localisation. C'est la proximité géographique des agrégats qui conduit à se tromper sur le modèle de choix de localisation. À l'inverse, quand les agrégats sont éloignés les uns des autres, même s'il est plus facile de retrouver un déterminant significatif, il est possible qu'un non déterminant soit trouvé significatif et défavorable à la localisation d'un établissement. Aussi, sur données empiriques, la coagglomération de plusieurs déterminants *a priori* peut conduire à se tromper sur les vrais déterminants de la localisation. La corrélation spatiale (qu'elle soit positive ou négative) des déterminants de la localisation testés devraient ainsi être pris en compte.

Paramètres associés aux secteurs

Chaque secteur d'inputs se voit assigné un paramètre qui définit son importance relative : 0, 1/3, 2/3 ou 1. Un secteur est dit déterminant si le paramètre est non nul. Nous vérifions si un déterminant est plus facilement identifié lorsque la valeur du paramètre qui lui est associé est plus élevée. Avant cela, nous vérifions si le modèle de localisation est mieux retrouvé lorsqu'il y a moins de secteurs déterminants. Le Tableau 3.4 donne la part des estimations conformes s'il y a un, deux ou trois secteurs déterminants. Le modèle estimé est plus conforme au modèle de localisation si le nombre de déterminants considérés est moins important. On retrouve à nouveau que le logit conditionnel est le modèle le moins performant. Si l'existence de facteurs multiples à la localisation conduit logiquement à réduire la part des simulations bien retrouvées dans les estimations, l'impossibilité de restituer le modèle de choix de localisation lorsque tous les facteurs possibles sont déterminants pose question. Quand trois secteurs sont déterminants le coefficient associé à chaque secteur est égal 1/3, les profits réalisables en chaque point j du territoire ne sont pas assez différenciés. Cela conduit à une répartition homogène des établissements sur le territoire et à ce qu'au moins un déterminant sur les trois ne soit pas significatif.

Le Tableau 3.5 montre bien que l'ensemble des spécifications permet de mieux retrouver un

secteur comme déterminant quand le paramètre associé dans le modèle de localisation est élevé.

Néanmoins, ces résultats peuvent aussi correspondre au fait qu'il n'y a qu'un seul secteur déterminant dans le modèle de localisation lorsque le coefficient associé est de 1 dans le modèle de choix de localisation.

Pour vérifier si le poids relatif du secteur par rapport aux autres a une influence sur les résultats, nous avons considéré l'échantillon des secteurs dont le paramètre était de 1/3 (Tableau 3.6). Dans ce cas, il peut y avoir un autre secteur déterminant dans le modèle de localisation (avec un paramètre de 2/3) ou deux autres secteurs déterminants (avec un paramètre de 1/3 pour chacun des deux secteurs). Quelle que soit la spécification utilisée, les secteurs déterminants sont mieux retrouvés quand les paramètres des autres secteurs significatifs sont plus faibles : plus un secteur est important dans le modèle de localisation, plus l'impact des autres secteurs sera négligeable et plus il sera facile d'en détecter la significativité.

Tableau 3.5 – Nombre de cas où un déterminant est trouvé comme significatif et de bon signe selon le paramètre du secteur

	1	2/3	1/3
Nombre de cas	24	48	72
MCD			
Logit	24	46	41
Probit	24	46	41
Logit conditionnel	24	44	39
MDC			
Poisson	24	46	41
Binomial négatif	24	46	41

Tableau 3.6 – Part des cas où un déterminant dont le paramètre est de 1/3 est trouvé comme significatif selon le paramètre des autres secteurs

	2/3	1/3
Nombre de cas	48	24
MCD		
Logit	26	15
Probit	26	15
Logit conditionnel	26	13
MDC		
Poisson	26	15
Binomial négatif	26	15

Variation du nombre de régions

À partir des données géographiquement agrégées selon la méthode de Voronoï, nous pouvons identifier la sensibilité des résultats au nombre de régions qui découpent le territoire. Pour cela nous

analysons les résultats obtenus à partir du modèle de Poisson. Nous nous contentons d'une analyse graphique montrant le nombre d'erreurs (Tableau 3.5), le nombre de déterminants (Tableau 3.6) et non déterminants (Tableau 3.7) correctement trouvés par les estimations. On constate globalement une amélioration des résultats au fur et à mesure que l'on augmente le nombre de régions découpant le territoire. En effet, lorsque le nombre de régions augmente :

— La performance n'est que faiblement améliorée lorsque l'on divise le territoire en plus de 13 régions. Ce nombre est dépendant de nos paramètres (nombre de secteurs, d'établissements,...), mais ce résultat met en lumière qu'il vaut mieux avoir des données issues d'un découpage géographique plus fin. Néanmoins, les résultats qualitatifs peuvent faiblement varier partir d'un certain nombre de régions, ce nombre étant propre à chaque problématique.

— Le nombre d'erreurs diminue car les informations sur chaque zone sont plus précises.

— Les déterminants sont mieux retrouvés lorsque le nombre de régions augmente (Tableau 3.6).

— Les non déterminants ne sont que faiblement impactés par l'augmentation du nombre de régions (Tableau 3.7.

FIGURE 3.5 – Variation du nombre d'erreurs par estimation selon le nombre de régions (Poisson)

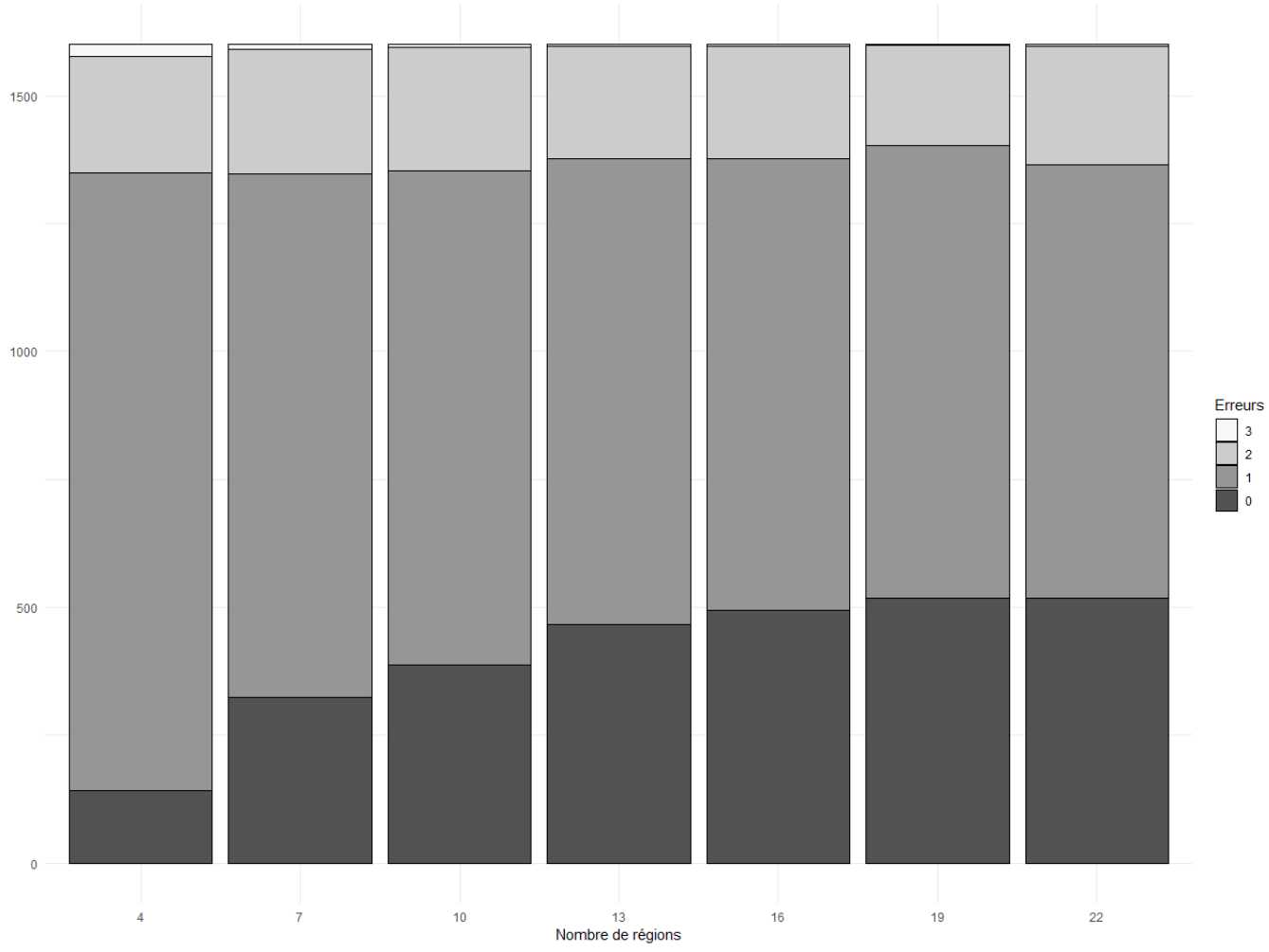
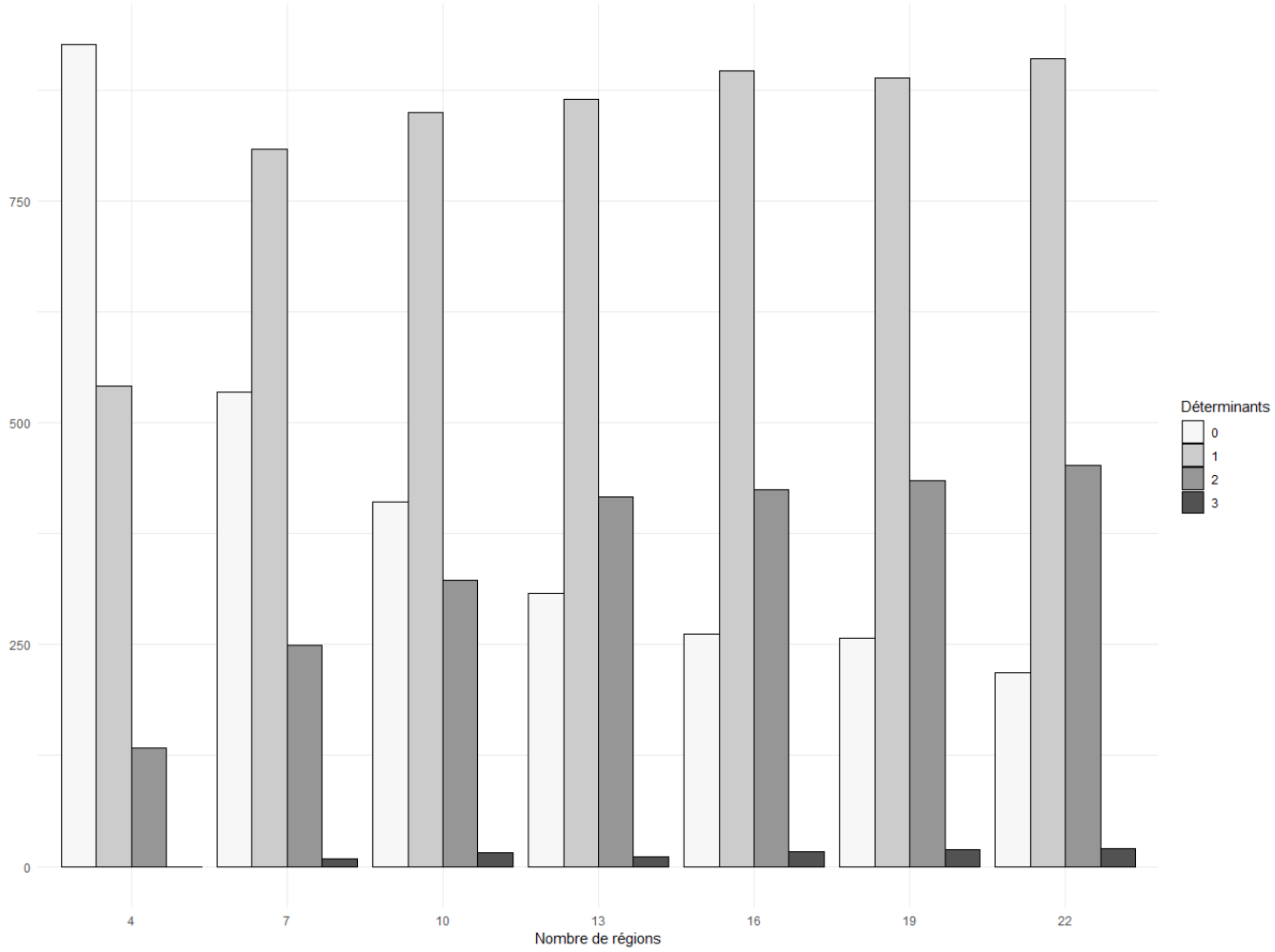


FIGURE 3.6 – Variation du nombre de déterminants conformément retrouvés selon le nombre de régions (Poisson)



3.4.2 Changement de processus de localisation

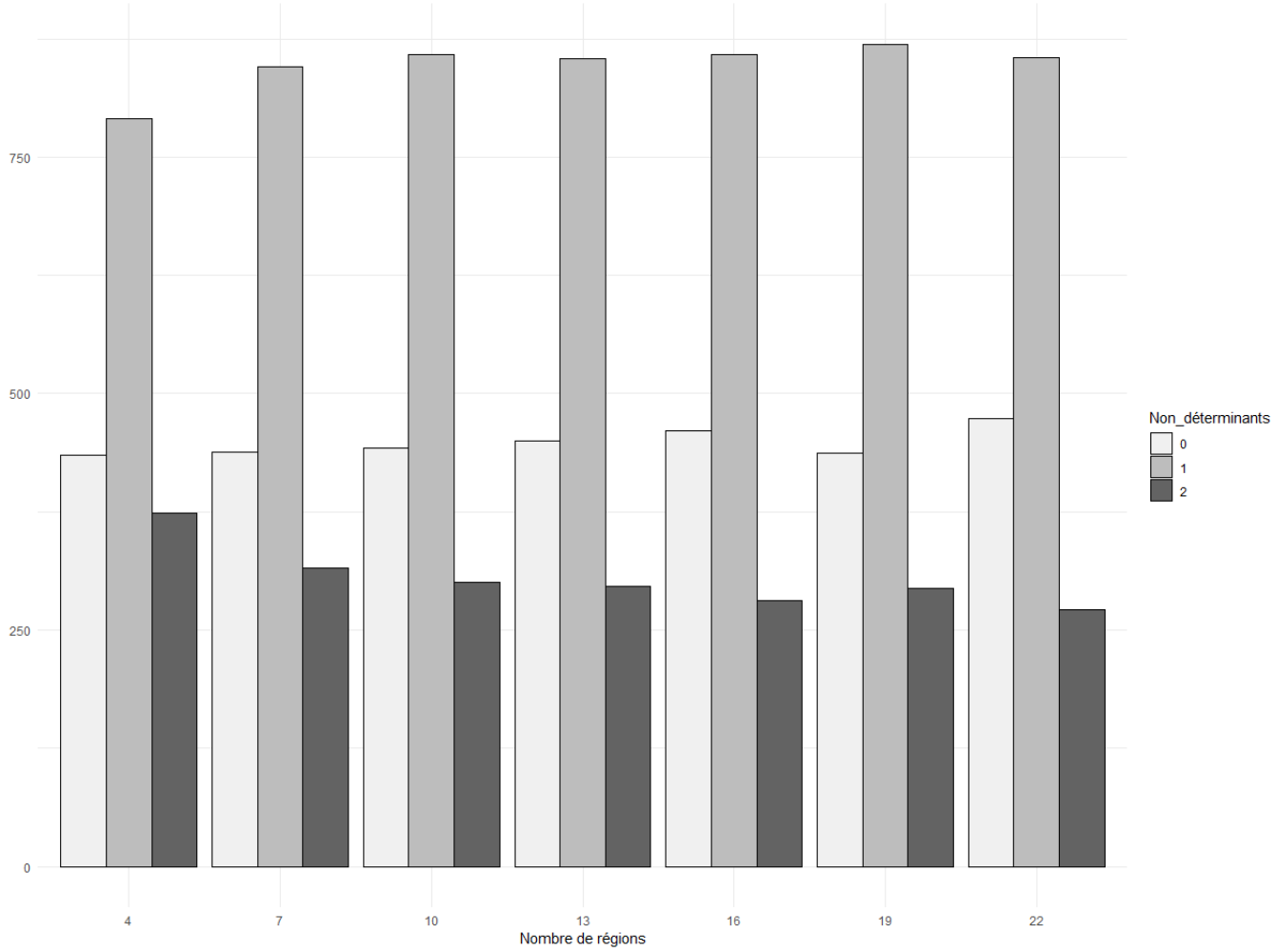
Nous avons également mené un travail similaire en modifiant les logiques de localisation possibles. Au lieu de considérer un processus aléatoire homogène et un processus de Matérn, nous prenons en compte un processus aléatoire de Poisson et un processus de Matérn. Cette modification implique donc qu'il n'y a plus nécessairement des établissements sur l'ensemble du territoire.

Les principaux résultats sont résumés par les points ci-dessous [5] :

— Le modèle de localisation initial est plus difficilement retrouvé. Plus particulièrement, les déterminants sont plus souvent significatifs, mais les non déterminants sont aussi plus souvent significatifs par rapport au premier modèle de simulation de localisation. Répartir des localisations selon un processus de Poisson conduit plus souvent à conclure qu'un facteur favorise la localisation, que celui-ci soit effectif ou non. On aura donc tendance à considérer comme déterminants des facteurs qui ne le sont pas nécessairement.

5. Les Tableaux sont en annexes

FIGURE 3.7 – Variation du nombre de non déterminants conformément retrouvés selon le nombre de régions (Poisson)



— Les résultats issus du modèle probit se distinguent du modèle logit, les résultats du probit sont plus performants. De plus, les estimations du logit conditionnel deviennent plus performantes que pour les autres estimations. Cela est imputable au fait que le logit conditionnel ne prend qu'un échantillon des observations, cela atténue le fait que les localisations ne soient plus réparties de manière homogène sur le territoire.

— L'augmentation du nombre de secteurs agglomérés diminue de manière monotone la qualité des résultats, on retrouve à nouveau le problème de la corrélation spatiale des localisations. Ainsi le changement du processus de localisation ne permet pas de supprimer ce problème, mettant d'autant plus en évidence l'importance d'en vérifier l'existence sur données empiriques.

— Les résultats issus des données agrégées sont globalement moins bien retrouvés.

3.5 Conclusion

Ce travail a pour but de comparer différentes spécifications économétriques généralement utilisées dans la littérature pour évaluer les facteurs de localisation des entreprises. Nos résultats suggèrent qu'il est difficile de savoir si les études empiriques menées sur données réelles permettent d'identifier avec pertinence la réalité des déterminants des choix de localisation. Pour étudier si l'une de ces spécifications peut être envisagée comme plus robuste que les autres, nous avons défini au préalable un modèle de localisation permettant ensuite de vérifier si les estimations de ces différentes spécifications sont conformes à notre modèle. Le choix d'un modèle théorique de localisation weberien (Weber, 1929) offre la possibilité d'implanter des établissements à partir de la localisation précise des facteurs et non de caractéristiques issues de découpage géographique. Nous proposons une méthode permettant d'implanter concrètement des établissements compte tenu de différents paramètres (localisation des autres secteurs, etc) à partir de spécifications ne reposant pas sur un découpage déterminé.

Nous observons que la logique de localisation des secteurs d'inputs (agglomérée ou aléatoire) peut fortement influencer la conformité des résultats d'estimations au modèle initial. D'autre part, l'importance relative de chacun des secteurs joue également sur cette conformité. Il n'y a aucune difficulté à identifier le secteur déterminant lorsque celui-ci est unique. En revanche, lorsqu'on a plus d'un secteur déterminant, les résultats sont meilleurs si tous les secteurs sont représentés. Nous avons également mis en évidence que la corrélation spatiale entre les agrégats d'établissements de différents secteurs pose problème pour identifier les déterminants du choix de localisation. De plus, l'application de découpages géographiques à différentes échelles montre qu'il est préférable d'utiliser des données désagrégées. À partir d'un certain seuil (qui dépend des autres paramètres du modèle) les résultats sont sensiblement identiques lorsque le nombre de régions découpant le territoire augmente.

Dans leur revue des travaux empiriques sur les modèles empiriques de choix de localisation, (Arauzo-Carod *et al.*, 2010) mentionnent une certaine défiance à l'égard de l'utilisation des MCD et plus particulièrement pour le logit conditionnel dans la mesure où celui-ci repose sur l'hypothèse relativement restrictive de l'indépendance des alternatives non pertinentes notamment. De plus, l'accessibilité des données à un niveau de désagrégation de plus en plus favorise l'utilisation des MDC. Les résultats issus de notre étude de conformité des modèles estimés sur données simulées à partir d'un modèle contrôlé semblent suggérer qu'à l'exception du logit conditionnel les deux familles de spécification (MCD et MDC) conduisent à des conclusions relativement proches sur l'impact des déterminants. Il faut néanmoins rester prudent sur la prise en compte des caractéristiques des

localisations pour identifier les déterminants des choix de localisation.

Du point de vue des décideurs publics, nos résultats suggèrent que les prises de décisions doivent se baser sur des analyses menées sur des données non agrégées et, idéalement, de construire les bases de données nécessaires à la conduite de ces types de travaux.

Pour approfondir ce travail de recherche, il serait intéressant d'accompagner cette démarche en proposant d'autres modèles de localisation ou se focalisant plus sur un paramètre (alors que nous avons privilégié une démarche plus exhaustive et exploratoire). La difficulté de ce type d'approche est accrue quand plusieurs phénomènes peuvent interagir et réduire les effets d'un phénomène spécifique. À ceci s'ajoute le choix du modèle qui, idéalement, ne doit pas dépendre d'un découpage territorial et permettre la comparabilité des résultats. La poursuite de ce travail pourrait aussi s'orienter sur la recherche sur données empiriques du nombre de zones minimal à partir duquel des résultats obtenus à partir de données agrégées peuvent converger.

3.A Annexes

Tableau 3.A.1 – Nombre de cas conforme

	Sans erreur	Une erreur	Deux erreurs	Trois erreurs
MCD				
Logit	38	40	2	0
Probit	42	36	2	0
Logit conditionnel	40	38	2	0
MDC				
Poisson	38	40	2	0
Binomial négatif	38	40	2	0

Tableau 3.A.2 – Nombre de cas conforme

	Déterminant			Non déterminant		
	Significatif	Non significatif	Erreur de signe	Significatif et positif	Non significatif	Significatif et négatif
MCD						
Logit	96	48	0	1	92	3
Probit	100	44	0	1	92	3
Logit conditionnel	99	45	0	2	91	3
MDC						
Poisson	96	48	0	1	92	3
Binomial négatif	96	48	0	1	92	3

Tableau 3.A.3 – Part des estimations conformes selon le nombre de secteurs d'inputs concentrés (en %)

	Zéro secteur	Un secteur	Deux secteurs	Trois secteurs
Nombre de cas	10	30	30	10
MCD				
Logit	70.00	46.67	46.67	30.00
Probit	80.00	53.33	50.00	30.00
Logit conditionnel	80.00	50.00	50.00	20.00
MDC				
Poisson	70.00	46.67	46.67	30.00
Binomial négatif	70.00	46.67	46.67	30.00

Tableau 3.A.4 – Part des estimations conformes selon le nombre de secteurs déterminants (en %)

	1	2	3
Nombre de cas	24	48	8
MCD			
Logit	95.83	31.25	0.00
Probit	95.83	39.58	0.00
Logit conditionnel	95.83	35.42	0.00
MDC			
Poisson	95.83	31.25	0.00
Binomial négatif	95.83	31.25	0.00

Tableau 3.A.5 – Part des cas où un déterminant est trouvé comme significatif selon le paramètre du secteur

	1	$^2/_3$	$^1/_3$
Nombre de cas	24	48	72
MCD			
Logit	23	38	11
Probit	23	39	11
Logit conditionnel	23	39	11
MDC			
Poisson	23	38	11
Binomial négatif	23	38	11

Tableau 3.A.6 – Part des cas où un déterminant dont le paramètre est de 1/3 est trouvé comme significatif selon le paramètre des autres secteurs

	$^2/_3$	$^1/_3$
Nombre de cas	48	24
MCD		
Logit	50.00	45.83
Probit	56.25	45.83
Logit conditionnel	54.17	45.83
MDC		
Poisson	50.00	45.83
Binomial négatif	50.00	45.83

FIGURE 3.A.1 – Variation du nombre d'erreurs par estimation selon le nombre de régions (Poisson)

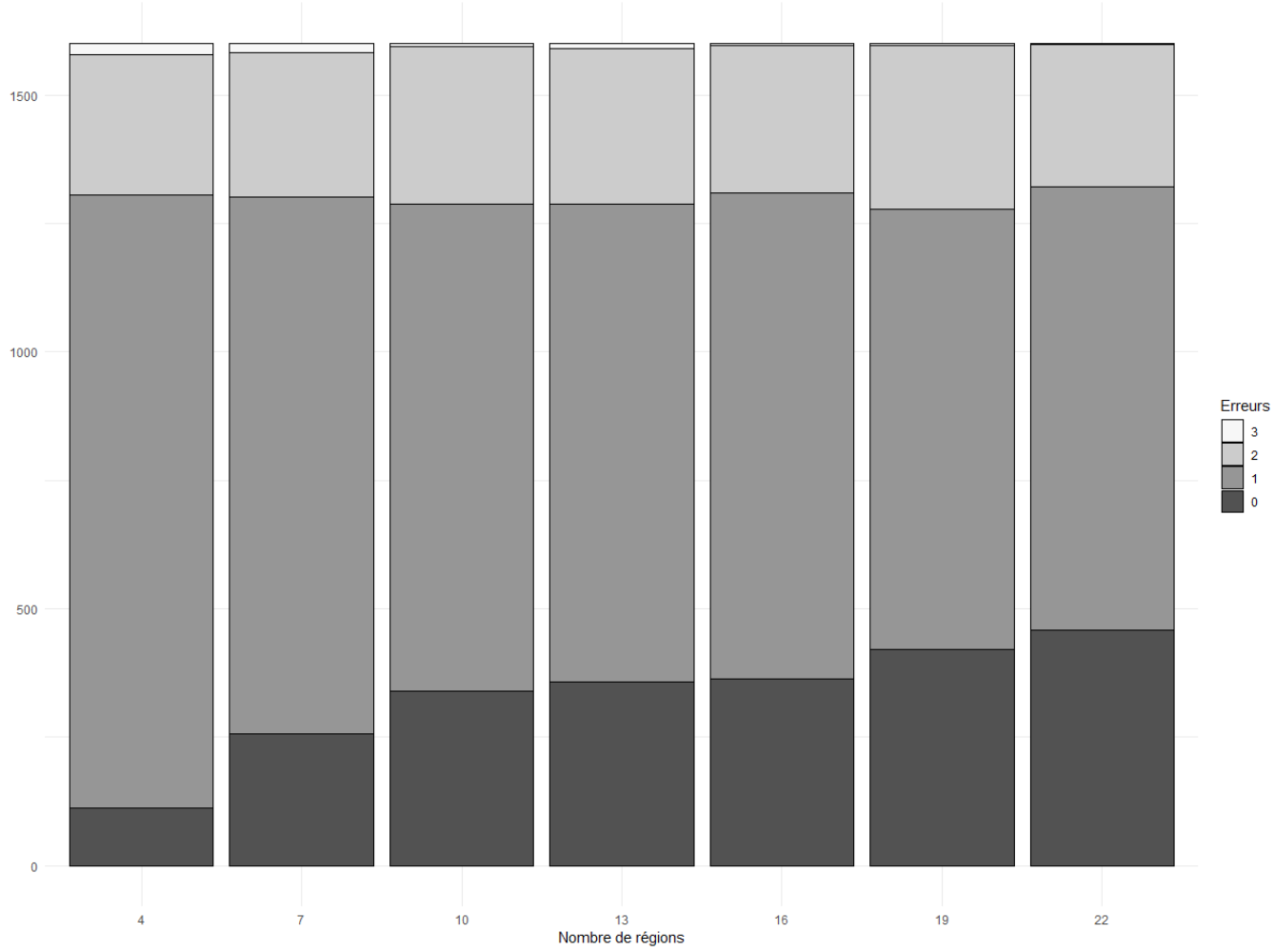


FIGURE 3.A.2 – Variation du nombre de déterminants conformément retrouvés selon le nombre de régions (Poisson)

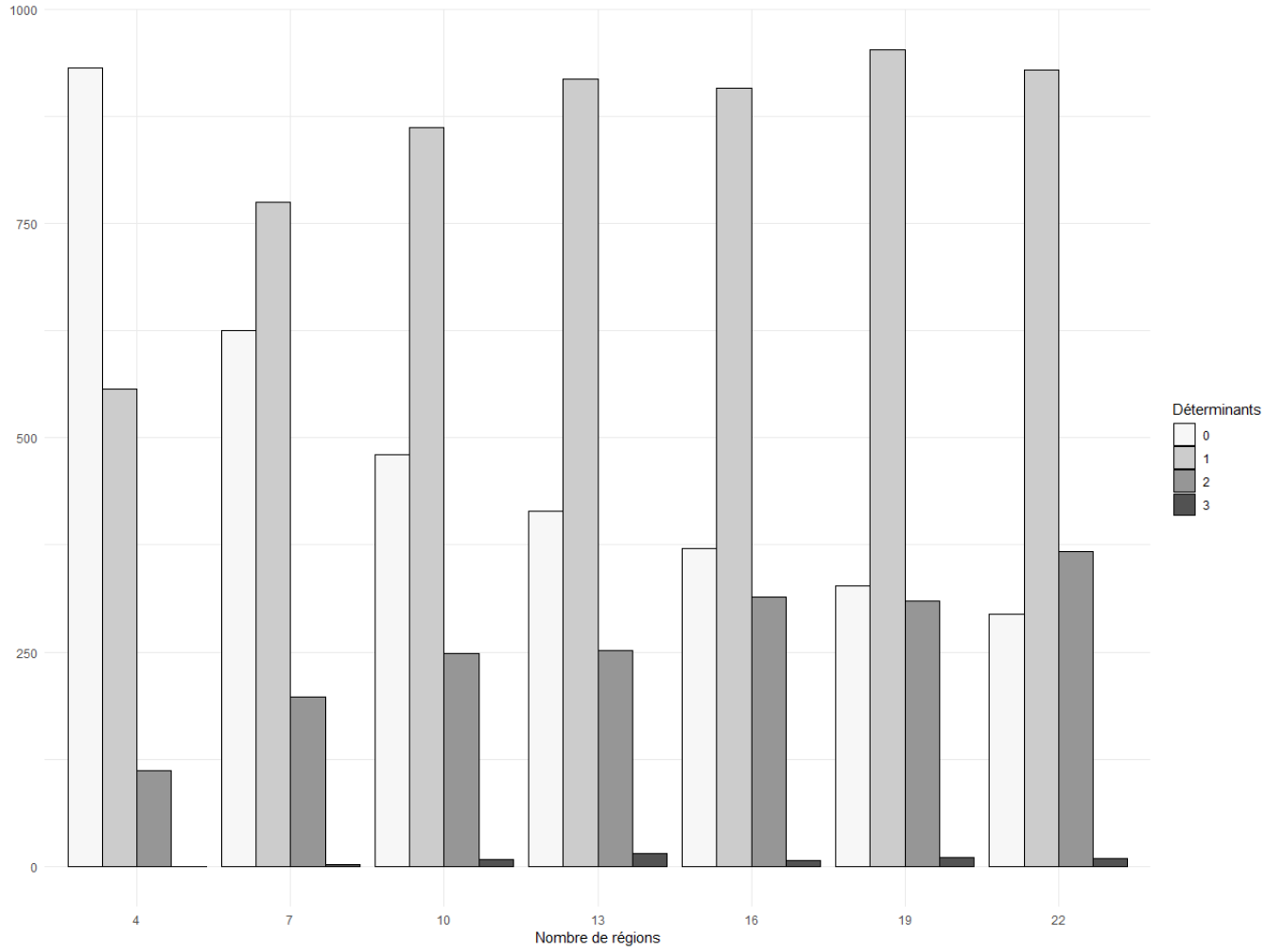
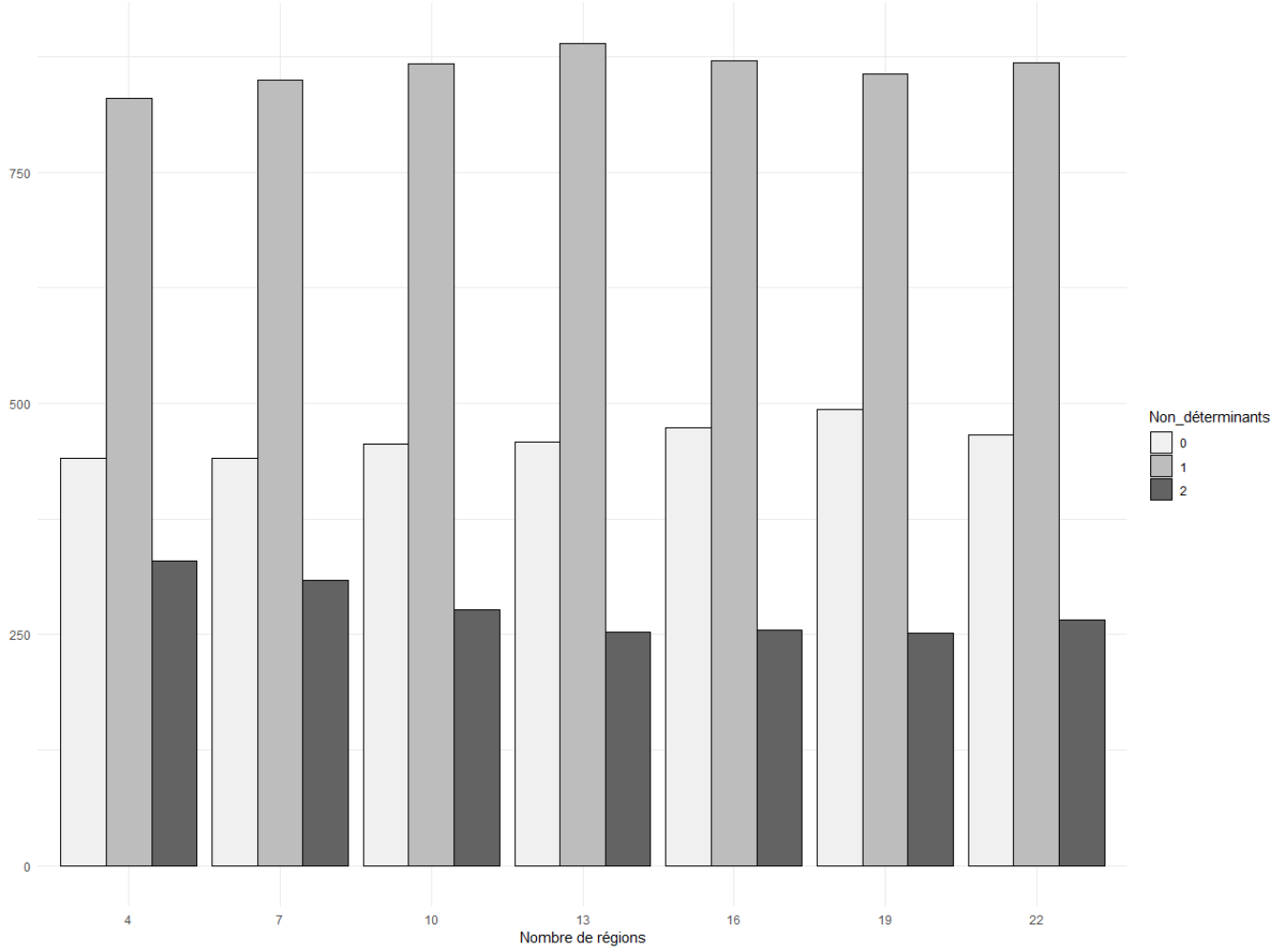


FIGURE 3.A.3 – Variation du nombre de non déterminants conformément retrouvés selon le nombre de régions (Poisson)



CHAPITRE 4

Objectiver le périmètre géographique pertinent pour une étude d'impact : les trajets quotidiens comme proxy

Ce chapitre fait l'objet d'un travail co-écrit avec Salima Bouayad Agha (GAINS)

Résumé

Un problème important des études d'impact (relatives aux manifestations sportives ou culturelles) réside dans la définition du périmètre géographique sur lequel il serait approprié de mesurer l'impact net. Si parfois le périmètre identifié semble pertinent, ce dernier n'est pas fondé sur un critère objectif. Ce travail propose un mode de détermination du périmètre approprié. L'un des éléments fondamentaux de ce projet est de s'affranchir des périmètres déjà existants. Puisque l'objectif principal dans les études d'impact est de parvenir à identifier les dépenses consenties par les spectateurs non locaux (les dépenses consenties par les spectateurs locaux ne doivent pas être prises en compte au motif que ceux-ci auraient également dépensé leur argent dans le bassin économique local en l'absence de la manifestation). Ce qui nous intéresse ici est d'identifier jusqu'à quelle limite géographique les consommateurs-travailleurs doivent être considérés comme appartenant au bassin économique local. Dans ce travail, nous proposons de nous fonder sur les trajets domicile-travail. Pour une commune donnée, celle où se déroule la manifestation sportive ou culturelle, seront considérés comme locaux les agents qui résident et travaillent dans la commune et ceux qui y travaillent dans la commune sans y résider (sous certaines conditions). C'est pourquoi l'étude des flux de mobilités professionnelles va nous aider à identifier les "bassins économiques locaux", c'est-à-dire les périmètres géographiques pertinents sur lesquels il sera possible de conduire les études d'impact. Dans ce travail nous proposons une méthode non dépendante du choix géographique, qui peut être utilisée de manière opérationnelle pour déterminer les territoires liés à un lieu de référence.

4.1 Introduction

L'organisation d'événements culturels ou sportifs est réputée stimuler l'activité d'un territoire. Les collectivités territoriales cherchent donc à mesurer, pour toute manifestation ou infrastructure existante ou projetée, quelles retombées économiques sont à attendre (Nicolas, 2007; Van Wyk *et al.*, 2015). L'une des raisons principales qui poussent les collectivités territoriales à mener des études d'impact est de déterminer si les manifestations ou les infrastructures engendrent des retombées économiques à la hauteur des investissements consentis. De plus, s'il existe un impact économique positif d'une manifestation pour un territoire voisin du lieu de l'événement, le porteur de la manifestation (une commune ou un département notamment) peut envisager de solliciter la collectivité voisine bénéficiaire pour qu'elle contribue au financement de l'événement.

Idéalement, l'organisation d'un événement doit créer un choc positif pour les agents économiques (entreprises, ménages) et doit perdurer dans le temps grâce à un effet multiplicateur : chaque euro injecté dans le périmètre géographique de l'événement doit permettre de créer plus d'un euro de richesse. En augmentant la demande de biens faite aux entreprises, la dépense envisagée ou effectuée est censée augmenter les bénéfices des entreprises et le niveau de salaire des travailleurs selon un cercle vertueux. Les pouvoirs publics qui supportent un coût d'organisation sont bénéficiaires sur le long terme. Pour des méga-événements, l'augmentation de la consommation (avant, pendant et après l'événement) entraine une hausse de certaines recettes fiscales (revenant à l'État). Lorsque ces événements sont plus modestes, ils peuvent permettre d'augmenter l'activité des entreprises locales et favoriser l'attractivité économique du territoire (aussi bien d'un point de vue productif que résidentiel, Sourd (2012)) [1].

Les organisateurs de manifestations sportives ou culturelles ont parfois tendance à ne pas anticiper les impacts négatifs que celles-ci peuvent engendrer. Les détracteurs de l'organisation de méga-événements invoquent le coût des infrastructures (les Jeux de Montréal 1976 qui ont engendré une dette que les contribuables locaux ont remboursé pendant 30 ans) et la difficulté à réutiliser celles-ci une fois l'événement terminé. Les opposants suggèrent également que les fonds destinés à l'organisation de ces événements auraient pu être orientés vers d'autres projets pouvant également augmenter les richesses et le bien être des habitants locaux. En conséquence, le choix de financer ces projets ne devrait pas se limiter à comparer les bénéfices espérés et les coûts de la manifestation, mais devrait également tenir compte de la possibilité de financer d'autres projets. Il ne faut pas non plus sous-estimer l'effet d'éviction lors des grands événements : des visiteurs potentiels renoncent

1. Hatem (2004) discute de la notion même d'attractivité, ainsi que des implications en terme de gouvernance des territoires et des causes de cette attractivité.

ou annulent leur visite en raison des désagréments liés à l'événement. L'effet d'éviction correspond également aux sommes dépensées par les locaux alors qu'elles auraient tout de même été faites pour d'autres événements locaux. Ces sommes ne doivent pas être considérées comme liées à l'événement car elles ne correspondent pas à une injection d'argent venue de l'extérieur (Baade, 2006; Baade et Matheson, 2016; Matheson, 2006). En effet, les injections des nouvelles dépenses sont celles des visiteurs ou spectateurs non locaux que l'on peut directement attribuer à l'événement, nettes des dépenses locales. Celles-ci ne doivent pas être prises en compte car les spectateurs locaux auraient consenti à des dépenses au sein du périmètre considéré même en l'absence de la manifestation envisagée. Dans ce contexte, le choix du périmètre adapté et du découpage géographique sous-jacent est crucial : mécaniquement le nombre de spectateurs extérieurs sera inversement proportionnel à la taille de la zone considérée et donc aux retombées économiques envisagées. Il est donc très important d'objectiver le périmètre géographique pertinent. De plus, le choix du périmètre pertinent doit permettre d'identifier le multiplicateur adapté à l'échelle géographique pertinente car la valeur du multiplicateur dépend de la structure productive du lieu d'accueil. En effet, l'effet multiplicateur n'est pas identique selon le secteur d'activité des établissements du territoire pris en compte (Crompton, 1995).

L'identification du périmètre adapté à partir d'un lieu de référence n'est pas spécifique aux événements sportifs ou culturel (Steiner *et al.*, 2015) On peut également y être confronté dans le cas de l'évaluation d'impact d'une politique publique. Par exemple, pour évaluer l'impact des politiques d'incitations aux entreprises dont l'objectif est de favoriser l'activité dans les Zones Franches Urbaines (ZFU). Mayer *et al.* (2015) trouvent un effet positif des politiques publiques pour une ZFU, mais constatent néanmoins un effet négatif sur l'activité des quartiers proches. Au global, l'effet de cette politique est nul car à une échelle géographique plus grande que celle de la ZFU il n'y a pas d'impact positif sur l'activité globale mais une redistribution des investissements entre les quartiers.

Pour répondre à cette question de la définition d'un périmètre géographique pertinent, nous proposons une stratégie d'identification du Bassin Économique Local (BEL) à partir d'un lieu cible en développant deux algorithmes spécifiques à ce travail. Comme il existe en fait autant de BEL que de localisations possibles pour un événement, il n'est évidemment pas possible de représenter tous les BEL du territoire en une seule carte. Notre approche repose sur l'utilisation des données mesurant les flux domicile-travail afin de cartographier un BEL particulier. Nous faisons l'hypothèse que pour une commune donnée, un visiteur est considéré comme local soit parce qu'il y réside, soit parce qu'il y travaille et que les déplacements entre sa commune de résidence et cette commune sont

nombreux. Ces flux servent donc de proxy pour mesurer l'attractivité économique des territoires ainsi que les facilités d'accès des réseaux de communications entre territoires (aussi bien la qualité que le coût pour les usagers).

Nous discuterons dans la prochaine section de la manière dont le périmètre pertinent est défini dans les études d'impact. Dans la section trois nous présenterons les différents découpages existants en France ainsi que la base de données mobilisée pour identifier les différentes catégories de territoires. Dans la section quatre, nous présenterons les deux algorithmes développés pour identifier les BEL, puis nous vérifierons si notre méthode permet de retrouver un découpage géographique déjà existant. Enfin, nous discuterons de l'apport et des limites de la méthode proposée.

4.2 Définition du périmètre géographique adapté pour une étude d'impact

La question de l'échelle géographique la plus adaptée pour étudier l'impact d'un événement est régulièrement évoquée dans la littérature (Crompton, 1995), mais il n'existe pas de critère permettant de l'objectiver. Par exemple, il n'existe pas de critère unanime pour apprécier si les JO ont un impact sur tout un pays ou juste les villes accueillant des compétitions. De même pour une coupe du monde de football : est-ce que tout le pays en profite, ou juste les villes hôtes, ou les départements avec une ville hôte, etc.

La zone géographique pertinente pour une étude d'impact dont on cherche à délimiter les contours n'a pas vocation à dépendre de la nature de l'événement ou de son importance. Quelle que soit la nature de l'événement organisé, les spectateurs locaux sont les mêmes. Il est donc important de distinguer le BEL que l'on cherche ici à caractériser, de ce que l'on définit communément comme une zone de chalandise. De manière générale, la zone de chalandise correspond à l'aire géographique d'où proviennent les clients d'un magasin. Cela veut donc dire que cette zone dépend de la localisation des magasins concurrents et du type de produits disponible (Croizean et Vyt, 2015). Si deux établissements n'appartenant pas au même secteur d'activité se localisent en un même point, la zone de chalandise ne sera pas la même. Dans le cas d'une étude d'impact, il est en revanche établi que la distinction des spectateurs en locaux/non locaux ne dépend pas du type d'événement. On ne peut pas définir le périmètre local en s'appuyant sur la définition d'une zone de chalandise. La concurrence d'événements substituables et géographiquement proches peut conduire à une segmentation territoriale des individus si ces derniers n'assistent qu'à un seul événement et non à plusieurs. Les individus assistent à l'événement le plus proche ce qui permet d'obtenir les zones de chalandise de chaque événement. Ainsi, les différents épreuves d'un championnat doivent permettre

à tous les spectateurs d'avoir un événement proche. Si les événements sont géographiquement concentrés, les prix des billets devront diminuer pour continuer à attirer des spectateurs sous peine de voir l'affluence diminuer. Pour autant, si les événements sont géographiquement proches, les spectateurs peuvent être considérés comme locaux pour chacun des événements.

Pour mener à bien une étude d'impact, trois critères peuvent être pris en compte pour identifier le périmètre pertinent (Maurence, 2012) :

— La distance kilométrique entre le domicile et l'événement : un spectateur est considéré comme local s'il réside à une distance inférieure à une distance seuil.

— Le temps de trajet entre le domicile et l'événement : un spectateur est considéré comme local si le trajet dure moins longtemps qu'une durée seuil.

— La localisation géographique du domicile et de l'événement : un spectateur est considéré comme local si le domicile et le lieu de l'événement appartiennent à la même zone géographique étant donné le découpage considéré (administratif ou socioéconomique).

Pour chacun de ces critères, le choix sous-jacent (une distance, une durée, un découpage géographique) influence l'étendue de la zone et conduit à sous-évaluer (ou sur-évaluer) le nombre de spectateurs locaux.

Généralement la prise en compte de la distance kilométrique et du temps de trajet entre deux localisations permet d'objectiver la proximité géographique de ces deux lieux. Ainsi, le fait d'utiliser le temps de trajet permet de corriger la distance kilométrique par la qualité du réseau routier. Le temps de trajet est donc plus adéquat pour étudier l'accessibilité entre deux territoires.

Ne considérer que l'appartenance géographique des individus conduit à une représentation limitée de la notion de proximité géographique entre deux lieux, et plus précisément de l'importance du lien entre lieu d'habitation et lieu de l'événement. Partir d'un découpage trop fin (c'est-à-dire avec un nombre de zones élevé), l'étude d'impact va sur-estimer le nombre de visiteurs non locaux.

À notre connaissance, la définition du périmètre local pour réaliser des études d'impacts recoupe celle des découpages géographiques administratifs (type département comme (Barget et Ferrand, 2012)) ou des zonages d'études (type bassin de vie). Les découpages administratifs reposent sur une répartition ancienne des activités. Les départements, par exemple, ont été créés en 1790 autour de chefs lieux qui depuis ont connu des trajectoires de développement contrastées. Les frontières "économiques" actuelles peuvent être très différentes de celles délimitées par les découpages administratifs. Pour cette raison, il existe aussi des découpages socio-économiques dont les périmètres ont été établis afin d'être en adéquation avec l'influence des plus grandes communes (selon les équipement nécessaires ou l'attractivité économique). On peut diviser ces découpages socio-économiques en deux sous-catégories, d'une part ceux qui incluent l'intégralité du territoire (Bassins de vie et

Zones d'emploi) et d'autre part ceux qui excluent les territoires non urbains (Aires urbaine et unités urbaines). Enfin, les EPCI (Établissement Public de Coopération Intercommunale) constituent un autre découpage pour lequel les territoires décident eux mêmes du groupe auquel se rattacher. Le détail de la constructions des zones non administratives est fourni en annexe à partir des définitions de l'INSEE [2] [3].

Néanmoins, l'appartenance à un territoire de deux communes ne signifie pas nécessairement que des liens directs existent entre les deux communes. De même, deux communes peuvent ne pas appartenir à une même zone géographique mais avoir des liens forts. Les découpages déjà existants ne sont pas non plus adéquats pour définir un périmètre adapté. Le Modifiable Areal Unit Problem (MAUP) analysé notamment par (Openshaw et Taylor, 1979, 1981) est ainsi présent si l'on utilise trop simplement les découpages existants. Même si certains découpages peuvent réduire l'influence du MAUP (les zonages non administratifs), tous y sont sensibles car l'existence de frontières implique nécessairement une dépendance au découpage choisi. La méthode proposée ici n'implique pas un découpage unique et représentable sur une seule carte pour l'ensemble des territoires mais elle repose sur une cartographie différente pour chaque territoire d'intérêt possible. Ainsi, un espace peut faire partie de plusieurs BEL.

De plus, le périmètre pertinent ne peut se restreindre aux frontières administratives délimitant le territoire du commanditaire de l'étude car cela risque de limiter l'impact à l'échelle du territoire du plutôt que de le prendre en compte sur le périmètre pertinent. Cela peut sur-évaluer les effets de l'événement en considérant à tort que des spectateurs sont des non locaux. Même si l'utilisation des découpages socioéconomiques est préférable aux découpages administratifs, elle reste délicate si le lieu de l'événement ne correspond pas au chef-lieu de la zone.

Les données de flux nous semblent plus adaptés pour évaluer les liens entre deux territoires. Ces liens peuvent être de différentes natures : les migrations domicile-travail, les migrations résidentielles, les flux établissement/siège, les flux transfert d'établissements. Nous faisons le choix d'utiliser les migrations domicile-travail car ces données reflètent l'attractivité des emplois des zones géographiques ainsi que la facilité de déplacement entre deux zones. De plus, les observations permettent de prendre en compte l'ensemble de la population active au lieu d'un simple échantillon (ce qui est le cas des migrations résidentielles) à l'échelle des individus (alors que les flux entre établissements ne permettent que d'avoir une information partielle du lien entre deux territoires).

Pour constituer le BEL à partir d'un lieu de référence (le lieu de la manifestation), nous agrégeons les territoires (communes ou autres) au sein desquels résident un grand nombre d'individus

2. Institut National de la Statistique et des Études Économiques.
3. La lecture d'Aliaga (2015) est à conseiller pour une description plus approfondie de ces divisions territoriales.

travaillant dans le territoire de référence. Nous faisons l'hypothèse qu'un individu ayant l'habitude de se déplacer vers le territoire de référence pour y travailler sera considéré comme local car il a l'habitude de s'y rendre, car il aura accès plus facilement aux informations concernant la manifestation (le lieu, l'heure,...) et car il consomme de nombreux biens et services commercialisés sur la commune du lieu de travail. Par construction, plus un territoire comptera d'habitants travaillant dans le territoire de référence et plus ce territoire aura de chance d'appartenir au BEL. Une fois le BEL délimité, un individu est considéré comme local s'il réside dans le BEL et un individu est considéré comme non local s'il réside pas dans le BEL.

4.3 Données

Afin de déterminer ce que nous considérons comme des BEL, nous partons de la base de données récapitulant les flux domicile-travail pour l'ensemble de la population. Ces mobilités sont des approximations intéressantes de l'attractivité économique des territoires (les emplois) et des moyens de transport reliant les territoires, que ce soit par la route, les transports en commun ou les deux à la fois. Elles permettent donc de rendre compte de l'importance du lien entre les territoires sans modélisation *a priori*.

Les données sont construites à partir du recensement de la population 2015 en prenant en compte le découpage géographique de 2017. En raison de l'étalement de la collecte des données en plusieurs vagues, un écart de comptabilité des flux peut exister mais sans que cela ne remette en cause les ordres de grandeur [4].

Pour ne nous restreindre qu'aux communes reliées par voie terrestres, nous ne considérons pas les flux entrants et sortants des départements et territoires d'Outre-mer, de la Corse, ainsi que des communes n'ayant aucun accès terrestre au continent (principalement des îles bretonnes).

Nous ne considérons pas non plus les flux de travailleurs vers l'étranger car nous ne disposons pas d'informations sur les flux de travailleurs venant de l'étranger pour travailler en France [5]. Le détail des destinations des travailleurs domiciliés en France mais travaillant à l'étranger (Tableau 4.1) indique un fort lien avec la Suisse et le Luxembourg. À l'inverse, à peine 6.5% de ces travailleurs ont comme destination l'Espagne ou l'Italie. Concernant les lieux de travail à l'étranger, nous trouvons la même importance des pays frontaliers que Floch (2011) sur les données du recensement 1999 et 2007 même si le nombre de navetteurs augmente au fil du temps (248 400 en 1999, 319

4. La documentation de l'INSEE fournit des explications complémentaires sur le recueil et l'utilisation possible de des données du recensement.

5. Nous aurions pu faire l'hypothèse que les travailleurs sont sédentaires (c'est-à-dire qu'ils vivent et travaillent dans la commune de résidence). Néanmoins, cela pourrait omettre la forte attractivité du territoire domestique ou du territoire étranger.

400 en 2007 et 404 710 en 2015). L'application de notre méthode pour des territoires frontaliers (plus particulièrement pour la Suisse) nécessite les données des navetteurs étrangers travaillant en France.

Tableau 4.1 – Pays des domiciliés français travaillant à l'étranger

Pays de destination	Nombre	Part des travailleurs
Suisse	186 629	46.1%
Luxembourg	76 343	18.9%
Allemagne	47 648	11.8%
Belgique	40 387	10.0%
Monaco	27 249	6.7%
Autre	26 454	6.5%
Somme	404 710	

Enfin, comme cela est rendu possible par les données nous divisons les grandes villes selon les arrondissements municipaux[6]. Chaque arrondissement est considéré comme une commune.

Le Tableau 4.2 présente les statistiques synthétiques sur les flux des travailleurs selon le découpage géographique utilisé. Par construction, le nombre moyen de travailleurs domicilié est égal au nombre moyen de travailleurs (domiciliés ou non) quel que soit le découpage pris en compte car il n'y a aucun individu entrant ou sortant du territoire national. On constate également une diminution des nombres moyens et médians quand le nombre de zones augmente, mais les valeurs de l'écart-type indique que la diminution du nombre de zones ne réduit pas l'hétérogénéité entre les zones. En effet, malgré un nombre de zones plus faible (et des valeurs moyennes et médianes également plus faibles), la dispersion du nombre de travailleurs domiciliés est plus élevée pour le découpage "Bassin de vie" que pour les découpages "Arrondissement" et "EPCI". On obtient également des valeurs médianes plus faibles pour le découpage "Zone d'emploi" que pour le découpage "Arrondissement". Cela s'explique de la manière suivante : les découpages non administratifs creusent l'écart entre les territoires attractifs et non attractifs. Les zones les plus importantes, qui incluent notamment Paris, Marseille et Lyon, constituent des observations qui augmentent considérablement la moyenne et l'écart type mais qui ont ont peu d'impact sur la médiane.

À partir des données de flux, nous construisons de nouvelles variables pour un découpage géographique donné :

— le nombre de travailleurs qui sont domiciliés dans la zone i ($Domiciliés_i$).

— le nombre de travailleurs qui travaillent dans la zone i, quel que soit le lieu de domiciliation ($Travailleurs_i$).

— le nombre de travailleurs domiciliés dans la zone i qui travaillent dans une autre zone géographique ($Sorties_i$).

— le nombre de travailleurs d'autres zones qui viennent travailler dans la zone i ($Entrées_i$).

6. 20 arrondissements pour Paris, 9 arrondissements pour Lyon et 16 arrondissements pour Marseille.

— le nombre de travailleurs qui vivent et travaillent dans la zone i ($Sédentaires_i$).

Ces informations nous permettent de définir trois ratios [7] qui synthétisent les flux pour les territoires :

— Le taux d'entrée d'un territoire mesure la part des salariés travaillant dans ce territoire et habitant en dehors de ce territoire :

$$\frac{Entrées_i}{Travailleurs_i} \times 100$$

— Le taux de sortie mesure la part des salariés résidant dans ce territoire et travaillant dans un autre territoire :

$$\frac{Sorties_i}{Domiciliés_i} \times 100$$

— Le solde relatif permet d'apprécier le solde des travailleurs entrants et sortants d'une zone par rapport au nombre de sédentaires :

$$\frac{Entrées_i - Sorties_i}{Sédentaires_i} \times 100$$

Le taux d'entrée peut s'interpréter comme l'apport de nouveaux travailleurs et donc de l'attractivité économique de la zone i (nombre d'emplois disponibles). À l'inverse, le taux de sortie correspond à la perte de travailleurs due à la faiblesse économique de la zone i. Intuitivement, cette attractivité économique correspond au nombre d'emplois disponible et aux salaires proposés. Néanmoins, des territoires peu attractifs économiquement voient une population y être domiciliée sans y travailler du fait des loyers faibles et de l'existence d'emplois nombreux à une distance géographique faible.

Tableau 4.2 – Flux de travailleurs

	Nombre de travailleurs domiciliés			Nombre de travailleurs			Nombre de travailleurs sédentaires		
	Moyenne	Médiane	Écart type	Moyenne	Médiane	Écart type	Moyenne	Médiane	Écart type
Région 2016	2 132 765	1 980 083	1 200 723	2 132 765	1 983 779	1 282 829	2 062 021	1 938 950	1 200 251
Région	1 218 723	831 600	1 104 222	1 218 723	817 901	1 166 087	1 171 975	799 498	1 097 711
Département	272 268	211 797	215 647	272 268	211 764	256 317	225 532	182 460	172 355
Zone d'emploi	86 172	40 031	186 753	86 172	37 549	224 219	69 130	30 810	161 175
Arrondissement	80 911	51 581	97 438	80 991	45 669	131 205	58 672	37 081	75 499
EPCI	20 892	9 208	98 153	20 892	6 762	119 967	14 103	4 162	88 677
Bassin de vie	15 701	5 794	125 411	15 701	4 297	138 803	11 490	2 522	121 998
Commune	732	187	3 411	732	71	4 973	248	40	1 926

Pour le découpage Commune, nous sélectionnons les communes avec au moins un travailleur domicilié, un travailleur qui travaille dans la commune, et un domicilié qui travaille dans la commune, soit 33 416 communes.

Les Tableaux 4.4 à 4.7 présentent, pour les découpages Région 2016, Région, Zone d'emploi et Bassin de vie, les zones avec les ratios les plus faibles et les plus élevés. Les Figures 4.1 à 4.4

7. Les taux d'entrée et de sortie rejoignent les définitions utilisées dans le document Insee (2011).

représentent le solde relatif pour les différents découpages [8].

Les résultats pour les différents découpages indiquent une forte attractivité des plus grandes agglomérations (Lyon, Marseille, Toulouse) car les taux d'entrée sont élevés et les taux de sortie faibles. D'autres territoires connaissent des taux de sortie et d'entrée faibles car les territoires ont des accès de communication faible et sont géographiquement isolés. Cet isolement peut être lié au fait qu'il s'agisse d'une zone de montagne (Clermont-Ferrand, Grenoble) ou à la proximité de la mer (Cherbourg, Nice, Toulon). Cela peut aussi être accentué par une faible activité économique du territoire et des territoires adjacents. En cas d'isolements géographique et économique, il est d'autant plus difficile pour le territoire de créer des emplois car le réseau de transport sera limité (forte congestion, qualité des transports, nombre de modes de transport possible). Par anticipation, les entreprises peuvent choisir des localisations alternatives pour ne pas se limiter à la main d'œuvre locale car le travailleur d'un territoire isolé doit vivre sur ce territoire (dans cette commune). Aussi, des spectateurs peuvent être considérés comme non locaux malgré une proximité géographique avérée.

Paris et les communes d'Île de France sont plus atypiques car les taux de sortie et d'entrée sont élevés. Ces zones sont attractives et entourées d'autres zones attractives. La proximité urbaines et des transports performants sont déterminants pour que ce type d'organisation puisse exister car les territoires ne doivent pas être isolés.

En regardant le solde relatif pour les différents découpages, nous pouvons affiner notre analyse. On constate par exemple la forte attractivité de la région Île de France au détriment de la région Hauts de France. En affinant le découpage, on constate en réalité que c'est l'ancienne région Picardie qui perd des travailleurs au détriment de la région Île de France, mais aussi de l'ancienne région Nord-Pas de Calais. L'agrégation due à la réforme régionale permet de bien mettre en lumière la sensibilité des conclusions possibles au découpage géographique. De manière similaire, nous constatons que ce sont les départements limitrophes de la région parisienne qui ont un solde de flux négatif.

Le Tableau 4.3 rend compte de l'effet du découpage géographique sur les ratios calculés. Nous vérifions bien que plus les zones sont agrégées, moins il y a de flux entre zones (le flux moyens des entrants est de 4.4% pour les régions 2016 contre 73.64% pour les communes). Le solde relatif nous indique également qu'il y a plus de zones déficitaires que de zones excédentaires car les valeurs médianes sont négatives. Les travailleurs vont donc majoritairement vers les mêmes territoires, là où les agglomérations économiques sont présentes, au détriment des territoires-dortoirs soumis aux

8. Les tableaux et les cartes sont en annexe pour les autres découpages français, à l'exception du découpe communal car les zones sont trop petites pour que la carte soit lisible

déplacements pendulaires des travailleurs. Ceci est également retrouvé comme un des déterminants de la mobilité des travailleurs par Lin *et al.* (2015). Les travaux indiquent clairement l'importance de l'emploi plutôt que du nombre d'habitants pour expliquer l'attractivité d'un territoire.

Nous avons vu que le choix du découpage géographique impacte l'analyse sur l'origine et la destination des navetteurs. Il n'est donc pas possible de se fier pleinement à un découpage existant pour établir le périmètre à utiliser pour évaluer l'impact d'un événement sportif ou culturel. Pour cela, nous développons un outil permettant d'objectiver ce périmètre tout en s'abstenant dans un découpage existant, et par conséquent des problèmes relevant du MAUP.

Tableau 4.3 – Flux de travailleurs

	Taux de sortie			Taux d'entrée			Solde relatif		
	Moyenne	Médiane	Écart type	Moyenne	Médiane	Écart type	Moyenne	Médiane	Écart type
Région 2016	4.04	3.14	2.43	2.97	2.43	1.39	-1.19	-1.04	2.96
Région	5.06	4.73	3.71	3.65	3.57	1.46	-1.68	-1.07	3.97
Département	14.08	9.84	11.48	11.59	8.53	10.54	-3.56	-1.76	15.64
Zone d'emploi	23.19	20.05	12.90	20.00	18.17	9.93	-7.97	-2.64	18.26
Arrondissement	27.06	24.03	14.41	21.99	19.56	10.94	-12.18	-7.27	26.15
EPCI	46.68	47.20	18.66	35.77	34.98	11.17	-54.42	-33.32	75.73
Bassin de vie	49.79	52.26	18.45	38.93	38.05	12.09	-58.31	-38.76	82.63
Commune	73.64	77.85	16.08	43.38	45.02	24.96	-304.07	-237.41	444.00

Données en pourcentage. Pour le découpage Commune, nous sélectionnons les communes avec au moins un travailleur domicilié, un travailleur qui travaille dans la commune et un domicilié qui travaille dans la commune, soit 33 416 communes.

Tableau 4.4 – Descriptifs des flux pour le découpage Région 2016

	Taux de sortie		Taux d'entrée		Solde relatif	
Ratio le plus élevé	9.36%	Centre-Val de Loire	6.38%	Île de France	5.55%	Île de France
	6.97%	Hauts de France	4.38%	Centre-Val de Loire	0.69%	Provence-Alpes-Côte d'Azur
	6.17%	Normandie	4.17%	Bourgogne-Franche-Comté	-0.02%	Auvergne-Rhône-Alpes
	5.25%	Bourgogne-Franche-Comté	3.12%	Normandie	-0.26%	Grand Est
	4.09%	Pays de la Loire	3.09%	Pays de la Loire	-0.82%	Nouvelle Aquitaine
	3.32%	Occitanie	2.70%	Provence-Alpes-Côte d'Azur	-1.00%	Bretagne
	2.96%	Bretagne	2.16%	Auvergne-Rhône-Alpes	-1.07%	Pays de la Loire
	2.78%	Nouvelle Aquitaine	2.01%	Bretagne	-1.18%	Bourgogne-Franche-Comté
	2.17%	Auvergne-Rhône-Alpes	2.00%	Nouvelle Aquitaine	-1.51%	Occitanie
	2.11%	Grand Est	1.89%	Hauts de France	-3.35%	Normandie
	2.04%	Provence-Alpes-Côte d'Azur	1.88%	Occitanie	-5.57%	Hauts de France
Ratio le moins élevé	1.24%	Île de France	1.86%	Grand Est	-5.75%	Centre-Val de Loire

12 zones

4.4 Détermination des frontières du BEL

Des méthodes existent pour cartographier les BEL à partir de flux, principalement des recherches sur les réseaux. Toussaint (1980) et Zhukov et Stewart (2013) offrent un panorama des Sphères d'Influence à partir de Graph (SIG). Ces méthodes de partitionnement consistent à diviser un réseau en catégorisant ses nœuds (ici les communes) en plusieurs sous-réseaux (ou "petits-mondes")

Tableau 4.5 – Descriptifs des flux pour le découpage Région

	Taux de sortie		Taux d'entrée		Solde relatif	
Ratio le	18.26%	Picardie	6.38%	Île de France	5.55%	Île de France
plus élevé	9.36%	Centre	6.36%	Bourgogne	1.31%	Alsace
	8.87%	Haute-Normandie	6.17%	Picardie	0.69%	Provence-Alpes-Côte d'Azur
	7.35%	Bourgogne	4.91%	Champagne-Ardenne	0.49%	Rhône-Alpes
	5.38%	Languedoc-Roussillon	4.76%	Limousin	0.03%	Limousin
	5.37%	Auvergne	4.49%	Basse-Normandie	-0.21%	Champagne-Ardenne
	5.11%	Basse-Normandie	4.38%	Centre	-0.21%	Midi-Pyrénées
	5.10%	Champagne-Ardenne	3.96%	Haute-Normandie	-0.68%	Basse Normandie
	4.91%	Franche-Comté	3.74%	Franche-Comté	-0.82%	Aquitaine
	4.73%	Limousin	3.63%	Poitou-Charentes	-1.00%	Bretagne
	4.73%	Poitou-Charentes	3.57%	Alsace	-1.07%	Pays de la Loire
	4.09%	Pays de la Loire	3.09%	Pays de la Loire	-1.14%	Bourgogne
	3.86%	Lorraine	3.04%	Auvergne	-1.19%	Poitou-Charentes
	2.96%	Bretagne	2.76%	Rhône-Alpes	-1.27%	Nord-Pas de Calais
	2.93%	Aquitaine	2.70%	Provence-Alpes-Côte d'Azur	-1.28%	Franche-Comté
	2.72%	Midi-Pyrénées	2.51%	Midi-Pyrénées	-1.71%	Lorraine
	2.56%	Nord-Pas de Calais	2.45%	Languedoc-Roussillon	-2.54%	Auvergne
	2.34%	Alsace	2.25%	Lorraine	-3.17%	Languedoc-Roussillon
	2.29%	Rhône-Alpes	2.15%	Aquitaine	-5.61%	Haute-Normandie
Ratio le	2.04%	Provence-Alpes-Côte d'Azur	2.01%	Bretagne	-5.75%	Centre
moins élevé	1.24%	Île de France	1.33%	Nord-Pas de Calais	-15.76%	Picardie

21 zones

Tableau 4.6 – Descriptifs des flux pour le découpage Zone d'emploi

	Taux de sortie		Taux d'entrée		Solde relatif	
5 Ratio les plus élevés	71.31%	Plaisir	62.49%	Plaisir	31.16%	Paris
	71.28%	Houdan	61.26%	Orly	27.71%	Ancenis
	62.80%	Orly	58.69%	Houdan	27.05%	Lille
	61.62%	Rambouillet	55.90%	Marne la Vallée	19.39%	Strasbourg
	61.05%	Poissy	47.53%	Poissy	18.01%	Honfleur
5 Ratio les moins élevés	6.24%	Aurillac	6.33%	Toulon	-75.35%	Mantes la Jolie
	6.46%	Briançon	6.33%	Gap	-74.12%	Étampes
	6.61%	Grenoble	6.40%	Grenoble	-72.18%	Rambouillet
	6.88%	Brest	6.63%	Cherbourg en Cotentin	-66.17%	Poissy
	6.98%	Clermont-Ferrand	6.81%	Aubenas	-57.78%	Dreux

297 zones

Tableau 4.7 – Descriptifs des flux pour le découpage Bassin de vie

	Taux de sortie		Taux d'entrée		Solde relatif	
5 Ratio les plus élevés	85.13%	Orry la Ville	88.93%	Saint Maximin	566.69%	Saint Maximin
	84.38%	Veyre-Monton	88.20%	Caudan	483.23%	Caudan
	83.72%	Marchiennes	87.94%	Ludres	403.14%	Ludres
	83.39%	Pont de l'Arche	85.65%	Crolles	379.81%	Vinon sur Verdon
	83.32%	Le Plessis-Bellevile	82.73%	Vinon sur Verdon	370.79%	Crolles
5 Ratio les moins élevés	6.26%	Toulouse	11.11%	Le Puy en Velay	-327.61%	Maintenon
	6.84%	Barcelonnette	11.40%	Paris	-324.98%	Saint Chéron
	6.94%	Briançon	11.79%	Puget Théniers	-319.85%	Esbly
	7.14%	Lyon	11.85%	Tende	-304.95%	Vimy
	7.24%	Bordeaux	11.92%	Die	-298.33%	Venerque

1630 zones

afin de réduire le nombre d'arcs du réseau. Les sous-réseaux ainsi créés ont ainsi des liens en intra élévés et des liens en inter faibles (Eusebio *et al.*, 2018).

Néanmoins les cartographies obtenues n'apportent qu'une information partielle car ces méthodes conduisent à rattacher une observation (un territoire) à un BEL. Or, un même territoire peut appartenir à plusieurs BEL. La Figure 4.5 représente le résultat du processus d'agrégation de marche aléatoire (Pons et Latapy, 2006), dont le but est d'identifier des groupes qui ont des flux

FIGURE 4.1 – Représentation du solde relatif pour le découpage Région 2016 (données en pourcentage)

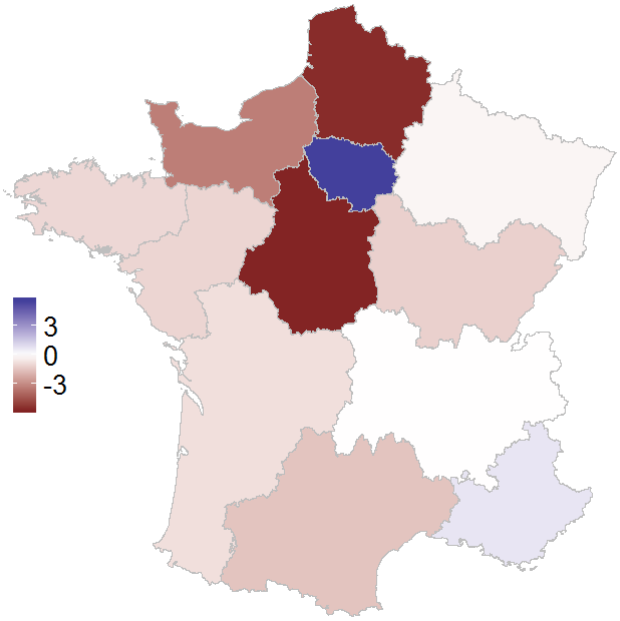
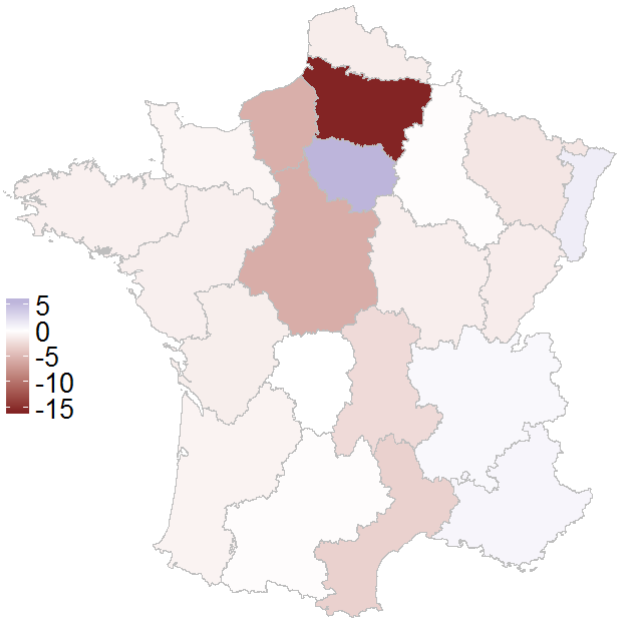


FIGURE 4.2 – Représentation du solde relatif pour le découpage Région (données en pourcentage)



intra-groupes élevés. Cela permet de retrouver les principales agglomérations mais il y a une surreprésentation de petites zones éloignées de ces agglomérations. Les zones sont plus petites le long de la diagonale des faibles densités (du Nord-Est au Sud-Ouest), car ces territoires souffrent de l'absence de pôles d'emploi attractifs (Oliveau et Doignon, 2016).

La méthode que nous proposons ne permet pas d'établir une cartographie complète d'un ter-

FIGURE 4.3 – Représentation du solde relatif pour le découpage Zone d'emploi (données en pourcentage)

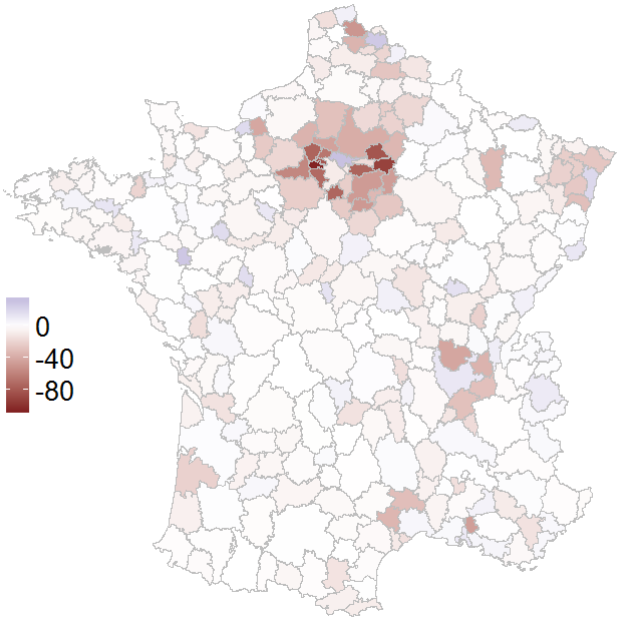
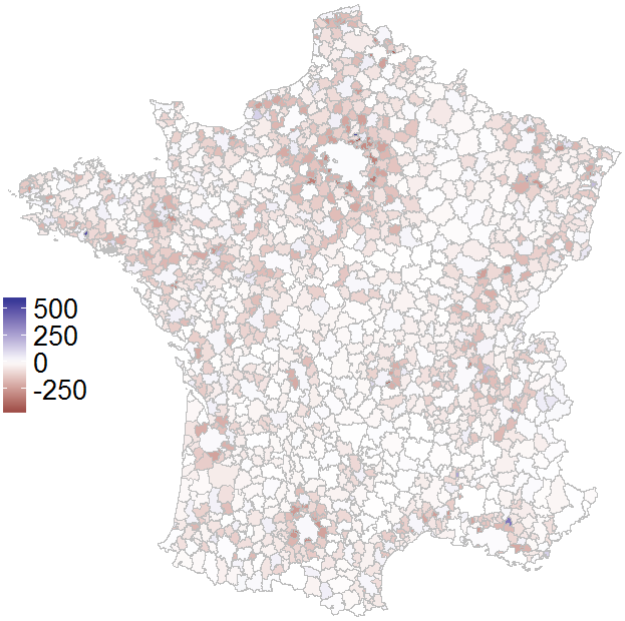
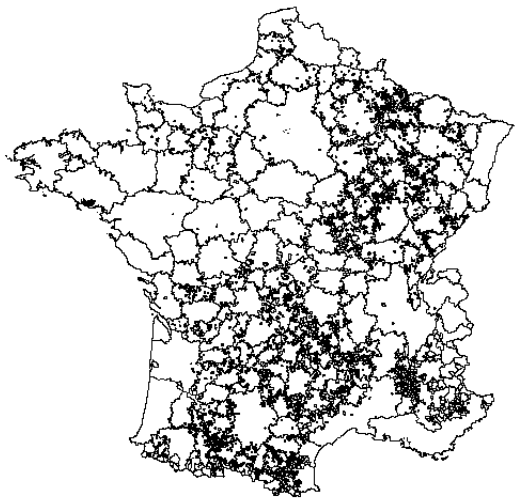


FIGURE 4.4 – Représentation du solde relatif pour le découpage Bassin de vie (données en pourcentage)



ritoire. En revanche, elle permet à partir d'un lieu référence d'identifier les territoires qui appartiennent à son BEL. Une fois celui-ci obtenu, on peut également vérifier la corrélation avec les découpages administratifs existants et juger de la pertinence d'un découpage géographique par rapport à son chef-lieu. Nous proposons deux algorithmes, ABSAPE (Algorithme Basé sur les

FIGURE 4.5 – Représentation de la France. Algorithme : Marche aléatoire.



Stocks Adjacents les plus Élevés) et ABAAPE (Algorithme Basé sur les Attractivités Adjacentes les plus Élevées).

Notre algorithme ABSAPE se décompose en plusieurs étapes :

1. Choix du lieu de référence (le lieu de l'événement). Cela peut être une commune ou un territoire plus grand. Le lieu de référence est le premier territoire faisant partie du BEL.
2. Sélection des lieux contigus au BEL. Cela peut également être une commune ou un territoire plus grand.
3. Sélection du lieu contigu le plus important, c'est-à-dire celui qui comprend le plus d'individus travaillant au lieu de référence. On l'agrège ensuite au BEL.
4. Vérification si un territoire n'est pas enclavé dans le BEL. Nous faisons l'hypothèse forte qu'un territoire fait partie d'un BEL s'il est enclavé [9]. Les BEL peuvent avoir des enclaves si ceux-ci sont formés de plusieurs territoires.
5. Retour à l'étape 2.
6. Le processus peut s'arrêter pour deux raisons. Soit il n'y a aucun flux de travailleurs entre

9. De manière symétrique, un territoire ne peut faire partie d'un BEL s'il n'existe pas un chemin avec le lieu de référence qui passe uniquement par ce BEL. Le BEL est supposé d'un seul tenant. Les hypothèses faites, qu'un territoire ne peut être enclavé et que la contiguïté doit être vérifiée entre les territoires appartenant à une même zone, ne sont pas toujours validées en réalité pour le territoire français. Même si ce n'est pas le seul exemple, on peut citer le cas du canton de Valréas, historiquement l'enclave des Papes. Ce canton fait partie de la région Provence-Alpes-Côte d'Azur mais est enclavé dans la région Auvergne-Rhône-Alpes. L'existence de territoires enclavés peut aussi être la conséquence de l'existence de communes composées de deux ou trois enclaves proches des zones frontalières pour un découpage donné, qu'il soit administratif ou socioéconomique. Le bassin de vie de Noisy-le-Roi est la seule exception car il est entièrement enclavé dans le bassin de vie de Paris. Il peut donc y avoir une justification économique à l'existence d'enclave. Nous autorisons l'existence d'enclaves de plusieurs territoires à l'intérieur d'un BEL.

le BEL et les zones contigües, soit un seuil est imposé pour se limiter aux flux les plus importants et le BEL créé dépasse ce seuil.

Nous proposons une variante de l'algorithme ABSAPE car en prenant en compte le stock de travailleurs entre territoires, cela risque de favoriser l'intégration rapide des territoires fortement peuplés au BEL, même si ces territoires sont éloignés. La différence intervient à l'étape 3. Notre variante ABAAPE consiste non plus à choisir le territoire avec le plus grand nombre de travailleurs qui viennent de ce territoire, mais le territoire qui, proportionnellement à son nombre de domiciliés, a le plus de travailleurs au lieu de référence du BEL. Cela permet d'agréger plus rapidement les territoires proches mais faiblement peuplés.

4.4.1 Application

Nous illustrons dans un premier temps la méthode proposée en choisissant comme lieu de référence la ville du Mans pour identifier les spectateurs locaux/non locaux des événements ayant lieu sur le circuit de la Sarthe. Notamment lors des 24h du Mans (de 200 000 à 250 000 spectateurs revendiqués depuis plus de vingt ans) et du Grand Prix moto (100 000 spectateurs en 2018) avec une présence visible de spectateurs étrangers mais aussi issus des départements limitrophes. Le Mans est chef lieu pour tous les découpages administratifs à l'exception des découpages régionaux (pour lesquels le chef lieu est Nantes) ce qui nous permet de réaliser une comparaison pertinente de notre découpage et de la quasi totalité des autres découpages. Enfin, les positions économique et géographique de la commune à l'échelle départementale sont centrales.

Les pôles les plus peuplés de l'agglomération mancelle sont les pôles urbains de Tours (77.87 kilomètres) et d'Angers (81.7 kilomètres). La ville du Mans n'est donc pas en concurrence directe d'une autre grande commune dans le sens où il n'existe pas un grand nombre de navetteurs du Mans à destination de ces deux pôles. La commune du Mans compte 144 244 habitants (25.35% de la population sarthoise) et l'unité urbaine mancelle 210 000 habitants (36.91% de la Sarthe) alors que la deuxième commune la plus peuplée, La Flèche, compte environ 15 000 habitants (2.64% de la population sarthoise). Il n'existe pas non plus de concurrence territoriale avec une autre ville du département, à la différence par exemple de Chartres et Dreux en Eure-et-Loir (respectivement 8.93% et 7.19% de la population départementale), de Rouen et du Havre en Seine-Maritime (13.74% et 8.79% de la population départementale).

Nous nous intéressons aussi à Alençon pour illustrer l'intérêt de notre méthode. Comme Le Mans, Alençon est une préfecture de département, le centre d'un bassin de vie, d'une zone d'emploi,... Outre le fait qu'Alençon est une commune plus petite que Le Mans dans un territoire moins peuplé que la Sarthe, Alençon présente l'intérêt d'être limitrophe de la Sarthe, et proche de la

frontière avec le département de la Mayenne (12 kilomètres). Elle est donc adéquate pour étudier le problème du MAUP lorsque l'on utilise des découpages administratifs.

Tableau 4.8 – Étalement des zones impliquant Alençon et Le Mans relativement au découpage départemental

	Zone d'emploi	Aire urbaine	EPCI	Bassin de vie	Unité urbaine
Superficie					
Alençon	71.21	56.81	73.32	49.82	70.59
Le Mans	98.13	100	100	100	100
Population					
Alençon	74.44	74.92	85.84	75.32	89.91
Le Mans	99.70	100	100	100	100

Données en pourcentage.

Le Tableau 4.8 illustre le fait que Le Mans a une place géographiquement centrale dans son département à la différence d'Alençon. Pour la première, c'est presque l'intégralité des populations et des superficies qui sont entièrement incluses dans le département. Pour Alençon en revanche, entre 74.44 et 89.91 % de la population sont dans le même département selon le découpage socioéconomique considéré. Si du point de vue de la population les valeurs peuvent paraitre élevées, la superficie dans le département n'excède pas 75%. C'est-à-dire que plus de la moitié de la superficie du bassin de vie d'Alençon se situe hors du département de l'Orne, dont Alençon est pourtant la préfecture. De manière concrète, cela peut poser des problèmes pour la ville d'Alençon de devoir prendre en compte un quart d'une population et la moitié d'une zone géographique qui ne sont pas dans le même département, et dans cet exemple, qui ne sont pas non plus dans la même région. Ces problématiques peuvent concerner les infrastructures de communication ou les établissements scolaires.

Pour l'instant les comparaisons ne peuvent s'effectuer qu'entre découpages existants. Ces découpages sont hétérogènes, donc le choix n'est pas anodin. Mais le fait de résumer un territoire par son appartenance à une entité est plus gênant car on ne connait pas l'influence des communes les plus importantes, les plus peuplées. À l'inverse, notre méthode permet de ne pas réduire une commune à son appartenance à une zone. Si notre méthode permet d'obtenir un BEL semblable à un découpage géographique, alors la prise de décision des décideurs sera d'autant plus simple car il ne sera pas nécessaire de procéder à des arbitrages pour prendre des décisions. En effet, si le BEL d'un lieu empiète sur plusieurs départements ou régions, les externalités territoriales seront à anticiper (positivement ou négativement) pour les autres départements ou régions.

4.4.2 Comparaison avec les découpages déjà existants

En premier lieu, il est important de noter que la valeur maximale n'est pas modifiée car les deux algorithmes s'arrêtent quand il n'y a plus de zones contiguës à agréger. La valeur maximale, et donc

FIGURE 4.6 – Représentation du BEL du Mans (Algorithme ABSAPE)

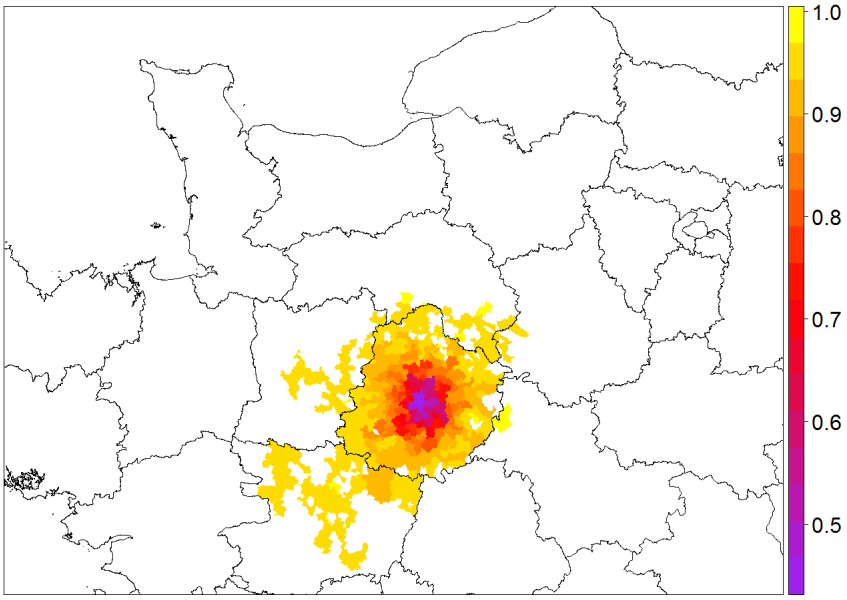


Tableau 4.9 – Algorithme ABSAPE-Commune : Le Mans

	Superficie (en km^2)	Population	Travailleurs domiciliés	Travailleurs au lieu de travail
Commune	52.75	147 121	52 326	81 432
75%	731.11	276 649	104 199	124 044
80%	973.98	301 198	115 176	128 005
90%	2 264.53	382 779	148 283	146 655
95%	5 303.55	732 262	276 204	300 174
Maximum	9 655.21	1 130 266	431 483	463 077

Commune au niveau de 46.64% et Maximum de 96.99%.

Tableau 4.10 – Algorithme ABSAPE-Commune : Le Mans. Origine de la population

Seuil	Région 2016	Région	Département	Zone d'emploi	Arrondissement	EPCI	Bassin de vie	Aire urbaine	Unité urbaine
75%	100	100	100	100	89.78	75.03	88.69	100	77.87
80%	100	100	100	100	87.14	69.72	88.24	100	71.53
90%	100	100	100	98.60	71.05	55.03	72.34	91.17	56.28
95%	100	100	71.49	56.61	37.14	28.76	37.88	48.55	29.42
Maximum	93.21	93.21	50.60	38.04	24.06	18.64	24.54	31.45	19.06

Commune au niveau de 46.64% et Maximum de 96.99%. Données en pourcentage.

Tableau 4.11 – Algorithme ABSAPE-Commune : Alençon. Origine de la population

Seuil	Région 2016	Région	Département	Zone d'emploi	Arrondissement	EPCI	Bassin de vie	Aire urbaine	Unité
75%	65.69	65.69	65.69	100	65.69	74.95	83.91	86.82	57.65
80%	67.16	67.16	67.16	99.09	67.16	66.70	74.71	78.74	49.95
90%	67.25	67.25	67.25	64.67	41.49	34.24	38.39	40.84	25.36
95%	42.85	42.85	42.85	26.66	18.94	13.31	14.92	15.87	9.85
Maximum	43.84	43.84	43.84	23.77	16.83	11.81	13.24	14.09	8.75

Commune au niveau de 34.04% et Maximum de 95.84%. Données en pourcentage.

le BEL maximal n'est impacté que si on impose un seuil minimum, en stock ou en part, en-dessous (ou au-dessus) duquel la zone ne serait pas à agréger. Comme les valeurs sont plus élevées pour l'algorithme ABAAPE, ce deuxième algorithme réduit l'effet du MAUP. En prenant en compte l'importance du flux relativement au poids des communes et non plus le nombre de travailleurs, on permet de réduire le poids des communes lointaines et fortement peuplées qui pouvaient très

FIGURE 4.7 – Représentation du BEL d'Alençon (Algorithme ABSAPE)

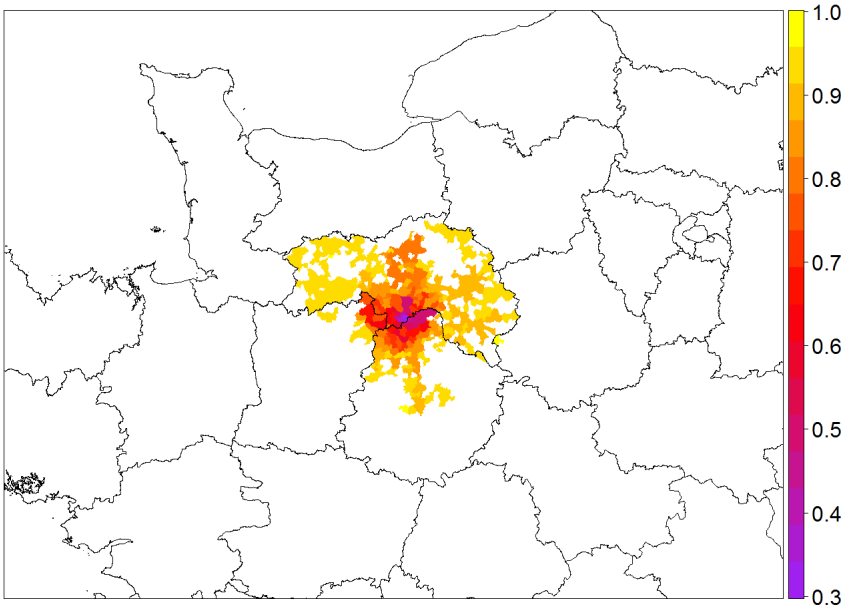


FIGURE 4.8 – Représentation du BEL du Mans (Algorithme ABAAPE)

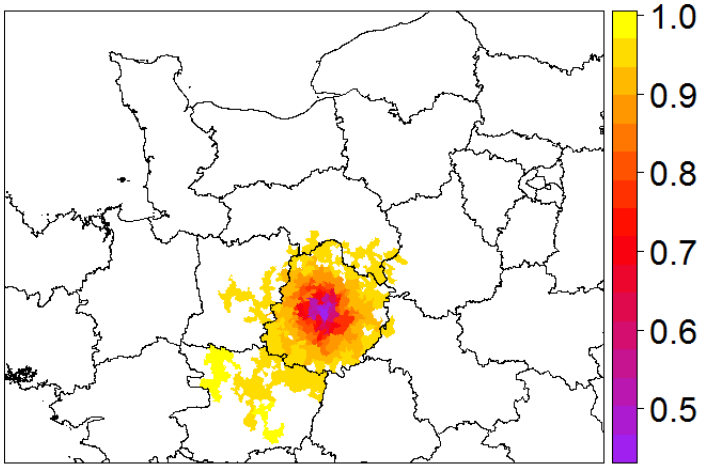
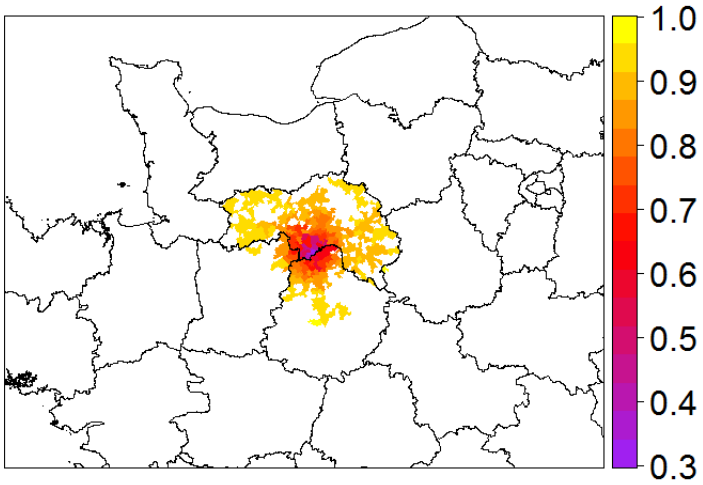


Tableau 4.12 – Algorithme ABAAPE-Commune : Le Mans. Origine de la population

Seuil	Région 2016	Région	Département	Zone d'emploi	Arrondissement	EPCI	Bassin de vie	Aire urbaine	Unité urbaine
75%	100	100	100	100	92.27	77.81	98.18	100	81.09
80%	100	100	100	100	89.37	70.48	92.11	100	73.45
90%	100	100	100	99.83	72.95	56.50	74.40	95.01	57.79
95%	100	100	99.34	78.62	50.12	38.82	51.11	65.51	39.70
Maximum	93.21	93.21	50.60	38.04	24.06	18.64	24.54	31.45	19.06

Commune au niveau de 46.64% et Maximum de 96.99%. Données en pourcentage.

FIGURE 4.9 – Représentation du BEL d'Alençon (Algorithme ABAAPE)



rapidement agrandir le BEL. Pour l'exemple du Mans, la valeur augmente entre 2 et 4 points de pourcentage pour les BEL au seuil de 75% et 80%.

Nous retrouvons des conclusions similaires en appliquant le deuxième algorithme à la ville d'Alençon. L'effet MAUP est accentué pour un seuil maximal car on intègre dans ce cas la commune du Mans.

Tableau 4.13 – Algorithme ABAAPE-Commune : Alençon. Origine de la population

Seuil	Région 2016	Région	Département	Zone d'emploi	Arrondissement	EPCI	Bassin de vie	Aire urbaine	Unité urbaine
75%	74.57	74.57	74.57	99.36	74.57	82.98	92.24	97.23	61.79
80%	71.65	71.65	71.65	98.89	71.65	75.23	84.36	89.45	55.73
90%	68.20	68.20	68.20	73.10	46.78	38.37	43.02	45.77	28.42
95%	72.67	72.67	72.67	41.02	27.93	20.38	22.85	24.31	15.09
Maximum	43.84	43.84	43.84	23.77	16.83	11.81	13.24	14.09	8.75

Commune au niveau de 34.04% et Maximum de 95.84%. Données en pourcentage.

4.4.3 BEL des circuits automobiles

Nous avons appliqué l'algorithme ABSAPE en prenant comme lieux de référence les circuits de Magny-Cours et du Castellet pouvant accueillir le Grand Prix de France de Formule 1 (Figures 4.10 et 4.11) [10]. Dans les deux cas, les territoires sont relativement isolés. Ces deux événements ont bénéficié du support des communes de Nevers pour Magny-Cours, et de l'action conjointe de Nice, Marseille et Toulon pour Le Castellet. Si Nevers et Magny-Cours sont dans leur BEL respectifs , ce

10. Nous avons également réalisé une cartographie du BEL de Carhaix-Plouguer, lieu du festival de musique "Les Vieilles Charrues". Comme pour Alençon une part importante de son BEL se situe hors de son département d'origine. La carte est en annexe.

n'est pas nécessairement le cas pour la deuxième localisation étudiée. Toulon est la seule des trois villes à faire partie du BEL du Castellet. De plus, Le Castellet ne fait pas partie du BEL de Nice, et fait partie de celui de Marseille uniquement si l'on prend en compte certains arrondissements. Ainsi, des territoires peuvent participer à l'organisation d'un événement qui n'a pas lieu sur leur BEL.

FIGURE 4.10 – Représentation du BEL de Magny-Cours (Algorithme ABSAPE)

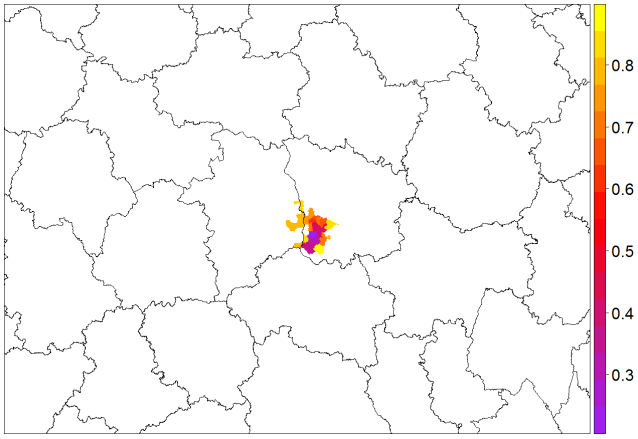
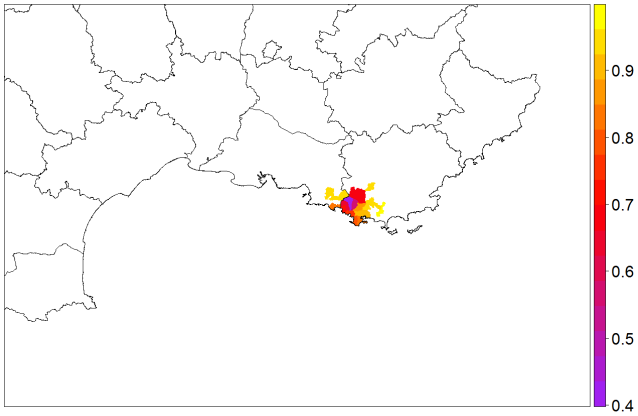


FIGURE 4.11 – Représentation du BEL du Castellet (Algorithme ABSAPE)



4.4.4 Corrélation des deux méthodes

Pour vérifier si les deux algorithmes donnent les mêmes BEL, nous calculons la corrélation entre ces deux méthodes. Pour ce faire nous prenons en compte les flux cumulés de travailleurs à chaque ajout d'une commune dans le BEL. Si la corrélation est égale à 1, cela veut dire que l'on intègre les communes au BEL dans le même ordre pour les deux algorithmes. Comme les deux méthodes partent de la même commune, que les mêmes communes font partie des BEL et que la contrainte de contiguïté limite les possibilités d'agrégation, les corrélations sont élevées et

significatives. Cependant, on peut comparer la corrélation des algorithmes selon le lieu d'intérêt. Le Tableau 4.14 donne les corrélations pour différents lieux.

Globalement, les corrélations de Pearson sont plus élevées que les corrélations de Spearman (sauf pour Le Mans), mais elles restent proches. L'ensemble des corrélations sont élevées à l'exception de Magny-Cours. Il serait intéressant de chercher, en utilisant plus de lieux de référence, les raisons d'une corrélation plus ou moins forte entre les deux algorithmes (isolement économique, flux faibles et peu nombreux,...).

Tableau 4.14 – Corrélation des algorithmes

Centre du BEL	Corrélation de Pearson	Corrélation de Spearman
Le Mans	0.87***	0.89***
Alençon	0.89***	0.85***
Paris (1)	0.85***	0.73***
Le Castellet	0.80***	0.74***
Magny-Cours	0.57**	0.47**
Carhaix-Plouguer	0.90***	0.84***

*** : corrélation significative à 1%. ** : corrélation significative à 5%.

4.5 Conclusion

La détermination du bassin économique local est une problématique récurrente dès lors que l'on veut évaluer une politique publique locale ou tout événement localisé. S'il y a erreur sur le périmètre du bassin, on peut sous-estimer ou sur-estimer l'impact des mesures prises par les décideurs publics. Les découpages utilisés peuvent être de différentes natures, administratives ou socioéconomiques.

L'utilisation de la base de données des flux domicile-travail de l'INSEE nous offre la possibilité d'avoir des données à une échelle fine (communale) et des données orientées entre les territoires. Ainsi, une commune peut être fortement reliée à une autre sans que cela ne soit réciproque.

Notre méthode autorise la création d'autant de BEL qu'il y a de lieux de référence possibles. Il n'est donc pas possible de résumer les découpages créés en une seule carte. Cela permet que nos résultats ne soient pas sensibles au MAUP car le territoire n'est pas découpé par des frontières fixes, mais par des frontières mouvantes qui dépendent du lieu de référence choisi.

Nos algorithmes permettent de définir une zone d'activités économiques pertinente et donc de connaitre l'origine de ces populations, et l'adéquation des découpages déjà existant à leur BEL qui n'est pas sensible au MAUP.

De multiples variantes de notre méthode peuvent être envisagées. Nous aurions par exemple pu décider de lier le BEL aux flux les plus importants selon un certain critère préalablement déterminés. Cette méthode sera fortement liée aux réseaux de communication (autoroutes, lignes

ferroviaires,...) et favoriserait ainsi les villes les plus accessibles et les plus peuplées. De plus cette méthode ne permet pas de constituer un BEL d'un seul tenant et sans enclave. Une autre variante serait de définir le voisinage non plus par la contiguïté mais selon un critère de distance ou, en s'appuyant sur les territoires contigus des territoires contigus du BEL, etc. Nous n'utilisons que les flux entre communes pour déterminer les BEL, mais la méthode proposée peut également appliquée pour des découpages plus agrégés. Il est également possible de prendre une zone agrégée, comme la région, et d'identifier le BEL en agrégeant des communes qui ne sont pas de la même région.

4.A Annexes

Zone d'emploi

Une zone d'emploi est un espace géographique à l'intérieur duquel la plupart des actifs résident et travaillent, et dans lequel les établissements peuvent trouver l'essentiel de la main d'œuvre nécessaire pour occuper les emplois offerts.

Le découpage en zones d'emploi constitue une partition du territoire adaptée aux études locales sur le marché du travail. Le zonage définit aussi des territoires pertinents pour les diagnostics locaux et peut guider la délimitation de territoires pour la mise en œuvre des politiques territoriales initiées par les pouvoirs publics ou les acteurs locaux. Ce zonage est défini à la fois pour la France métropolitaine et les DOM.

Le découpage actualisé se fonde sur les flux de déplacement domicile-travail des actifs observés lors du recensement de 2006.

EPCI

Les établissements publics de coopération intercommunale (EPCI) sont des regroupements de communes ayant pour objet l'élaboration "de projets communs de développement au sein de périmètres de solidarité ". Ils sont soumis à des règles communes, homogènes et comparables à celles de collectivités locales. Les communautés urbaines, communautés d'agglomération, communautés de communes, syndicats d'agglomération nouvelle, syndicats de communes et les syndicats mixtes sont des EPCI.

Bassin de vie

Le bassin de vie constitue le plus petit territoire sur lequel les habitants ont accès aux équipements et services les plus courants. On délimite ses contours en plusieurs étapes. On définit tout d'abord un pôle de services comme une commune ou unité urbaine disposant d'au moins 16 des 31 équipements intermédiaires. Les zones d'influence de chaque pôle de services sont ensuite délimitées en regroupant les communes les plus proches, la proximité se mesurant en temps de trajet, par la route à heure creuse. Ainsi, pour chaque commune et pour chaque équipement non présent sur la commune, on détermine la commune la plus proche proposant cet équipement. Les équipements intermédiaires mais aussi les équipements de proximité sont pris en compte.

La méthode ANABEL permet enfin d'agréger par itérations successives les communes et de dessiner le périmètre des bassins de vie comme le plus petit territoire sur lequel les habitants

ont accès aux équipements et services les plus courants. Le zonage en bassins de vie apporte un complément à travers l'analyse de la répartition des équipements et de leur accès.

Son principal intérêt est de décrire les espaces non fortement peuplés, c'est à dire les bassins de vie construits sur des unités urbaines de moins de 50 000 habitants.

Aire urbaine

Une aire urbaine ou "grande aire urbaine" est un ensemble de communes, d'un seul tenant et sans enclave, constitué par un pôle urbain (unité urbaine) de plus de 10 000 emplois, et par des communes rurales ou unités urbaines (couronne périurbaine) dont au moins 40% de la population résidente ayant un emploi travaille dans le pôle ou dans des communes attirées par celui-ci.

Le zonage en aires urbaines 2010 distingue également :

— les "moyennes aires", ensemble de communes, d'un seul tenant et sans enclave, constitué par un pôle (unité urbaine) de 5 000 à 10 000 emplois, et par des communes rurales ou unités urbaines dont au moins 40% de la population résidente ayant un emploi travaille dans le pôle ou dans des communes attirées par celui-ci.

— les "petites aires", ensemble de communes, d'un seul tenant et sans enclave, constitué par un pôle (unité urbaine) de 1 500 à 5 000 emplois, et par des communes rurales ou unités urbaines dont au moins 40% de la population résidente ayant un emploi travaille dans le pôle ou dans des communes attirées par celui-ci.

Les aires urbaines, datées de 2010, ont été établies en référence à la population connue au recensement de 2008.

Unité urbaine

L'unité urbaine est une commune ou un ensemble de communes qui comporte sur son territoire une zone bâtie d'au moins 2 000 habitants où aucune habitation n'est séparée de la plus proche de plus de 200 mètres. En outre, chaque commune concernée possède plus de la moitié de sa population dans cette zone bâtie.

Si l'unité urbaine s'étend sur plusieurs communes, l'ensemble de ces communes forme une agglomération multicommunale ou agglomération urbaine. Si l'unité urbaine s'étend sur une seule commune, elle est dénommée ville isolée.

Les unités urbaines, datées de 2010, ont été établies en référence à la population connue au recensement de 2007.

Descriptifs et cartographie des flux

Tableau 4.A.1 – Descriptifs des flux pour le découpage Département

	Taux de sortie		Taux d'entrée		Solde relatif	
10 Ratio les plus élevés	53.87%	Seine-Saint Denis	58.31%	Paris	95.33%	Paris
	53.55%	Val de Marne	58.21%	Hauts de Seine	50.49%	Hauts de Seine
	50.00%	Val d'Oise	49.15%	Seine-Saint Denis	13.15%	Rhône
	47.04%	Hauts de Seine	45.85%	Val de Marne	8.21%	Doubs
	42.89%	Essonne	39.08%	Val d'Oise	6.04%	Drôme
	42.60%	Seine et Marne	31.38%	Yvelines	5.34%	Haute Garonne
	41.77%	Yvelines	27.72%	Essonne	4.40%	Bouches du Rhône
	30.73%	Paris	27.22%	Territoire de Belfort	3.88%	Nord
	29.34%	Oise	21.60%	Seine et Marne	3.44%	Haute Marne
	29.34%	Eure	18.71%	Drôme	3.23%	Vaucluse
10 Ratio les moins élevés	1.98%	Alpes maritimes	2.23%	Pyrénées orientales	-46.68%	Seine et Marne
	3.29%	Pyrénées orientales	3.41%	Var	-36.77%	Essonne
	3.41%	Gironde	3.47%	Alpes maritimes	-35.82%	Val d'Oise
	4.00%	Haute Savoie	3.82%	Finistère	-30.65%	Val de Marne
	4.50%	Bouches du Rhône	3.88%	Gironde	-28.25%	Oise
	4.52%	Bas-Rhin	4.30%	Ardennes	-26.21%	Eure
	4.73%	Finistère	4.63%	Cantal	-26.00%	Yvelines
	4.77%	Haute Garonne	4.81%	Aveyron	-24.11%	Eure et Loir
	4.86%	Hautes Alpes	5.09%	Puy de Dôme	-22.58%	Haute Saône
	5.06%	Puy de Dôme	5.16%	Charente maritime	-20.15%	Seine-Saint Denis

94 zones

Tableau 4.A.2 – Descriptifs des flux pour le découpage Arrondissement

	Taux de sortie		Taux d'entrée		Solde relatif	
5 Ratio les plus élevés	70.52%	Nogent sur Marne	71.50%	Boulogne Billancourt	99.99%	Nanterre
	67.43%	Antony	70.89%	Saint Denis	95.53%	Paris
	67.41%	Bobigny	70.59%	L'Haÿ les Roses	60.35%	Saint Denis
	66.98%	Argenteuil	66.84%	Bobigny	50.61%	Boulogne Billancourt
	66.95%	L'Haÿ les Roses	66.29%	Nanterre	43.08%	Versailles
5 Ratio les moins élevés	7.00%	Toulouse	7.83%	Millau	-87.52%	Meaux
	7.05%	Bordeaux	8.01%	Draguignan	-82.51%	Le Raincy
	7.20%	Limoges	8.43%	Thonon les Bains	-79.93%	Provins
	7.56%	Gap	8.54%	Toulon	-78.74%	Muret
	7.57%	Cherbourg	8.83%	Brest	-70.79%	Lodève

316 zones

Tableau 4.A.3 – Descriptifs des flux pour le découpage EPCI

	Taux de sortie		Taux d'entrée		Solde relatif	
5 Ratio les plus élevés	85.95%	CC Plaines et Monts de France	78.32%	CC Jalle-Eau-Bourde	169.06%	CA Val d'Europe Agglomération
	82.51%	CC Norge et Tille	78.29%	CC de l'Est Lyonnais	131.39%	CC Caux Estuaire
	82.46%	CC Vallées de l'Orne et de l'Odon	77.99%	CA Val d'Europe Agglomération	111.31%	CC de l'Est Lyonnais
	81.92%	CC des Coteaux Bellevue	77.95%	CC Caux Estuaire	92.40%	CC Rives de Moselle
	81.56%	CC Haut Chemin-Pays de Pange	75.66%	CC de la Vallée du Garon	84.08%	CC Jalle-Eau-Bourde
5 Ratio les moins élevés	9.01%	Métropole Européenne de Lille	12.40%	CA Lannion-Trégor Communauté	-330.88%	CC du Triangle Vert
	9.16%	CA du Bassin d'Aurillac	12.44%	CC Couserans-Pyrénées	-303.16%	CC des Coteaux Bellevue
	9.31%	Métropole du Grand Paris	12.63%	CA de Saint-Dié-Des-Vosges	-295.02%	CC de la Haute Deûle
	9.33%	CA du Pays Basque	12.72%	CC du Massif du Vercors	-293.98%	CC du Kochersberg
	9.47%	CC du Pays de Gex	12.85%	CC de Saint-Flour	-287.20%	CC de Seille et Mauchère et Grand Couronné

1225 zones

Tableau 4.A.4 – Descriptifs des flux pour le découpage Commune

	Taux de sortie		Taux d'entrée		Solde relatif	
5 Ratio les plus élevés	99.51%	Bihorel	99.18%	Bihorel	566.69%	Roissy en France
	90.64%	Sainte Catherine	98.70%	Roissy en France	483.23%	Rungis
	90.34%	Vaulx-Milieu	97.19%	Rungis	403.14%	Labège
	88.86%	Hallennes lez Haubourdin	96.07%	Labège	379.81%	Sochaux
	88.83%	Vezin le Coquet	95.84%	Chessy	370.79%	Chessy
5 Ratio les moins élevés	12.83%	Mende	29.72%	Biscarosse	-480.28%	Thue et Mue
	13.71%	Les Belleville	30.98%	Nice	-474.81%	Andrésy
	14.38%	Gap	31.65%	Marseille 12	-474.81%	Vauréal
	14.59%	Aurillac	32.38%	Sainte Maxime	-468.43%	Saint Germain lès Corbeil
	15.22%	Courchevel	32.63%	Marseille 7	-466.37%	Pibrac

34 956 zones ramenées à 2 999 pour avoir des communes avec au moins 1 000 travailleurs domiciliés et 1 000 travailleurs qui travaillent dans la commune.

FIGURE 4.A.1 – Représentation du solde relatif pour le découpage Département (données en pourcentage)

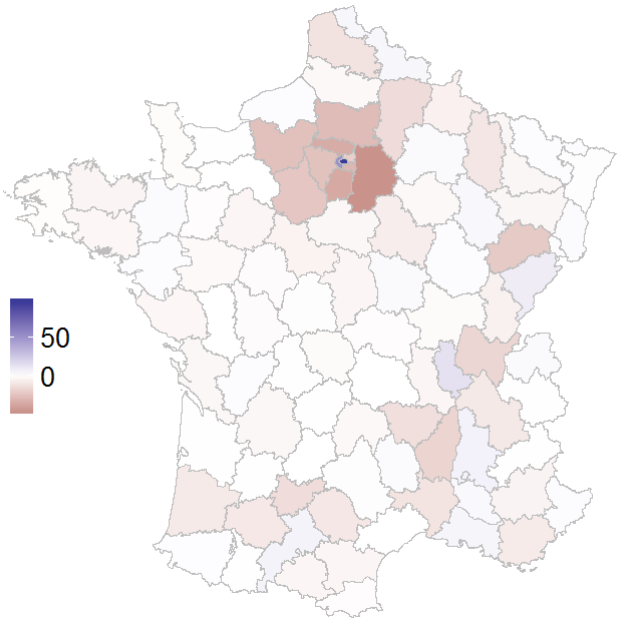


FIGURE 4.A.2 – Représentation du solde relatif pour le découpage Arrondissement (données en pourcentage)

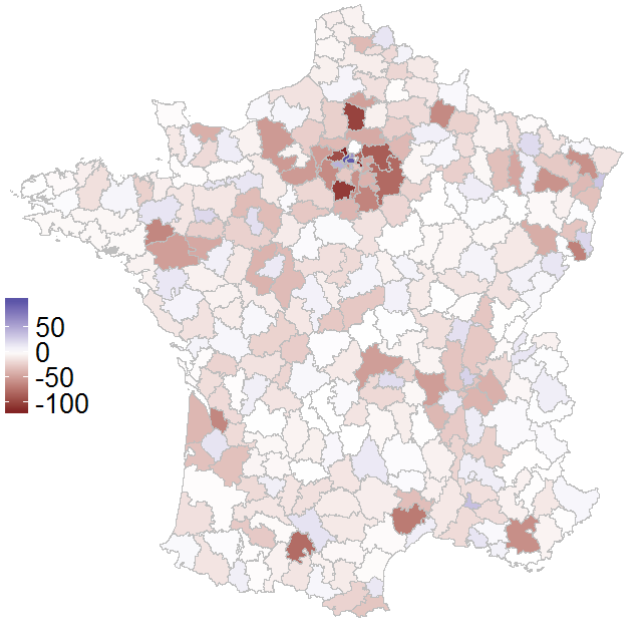


FIGURE 4.A.3 – Représentation du solde relatif pour le découpage EPCI (données en pourcentage)

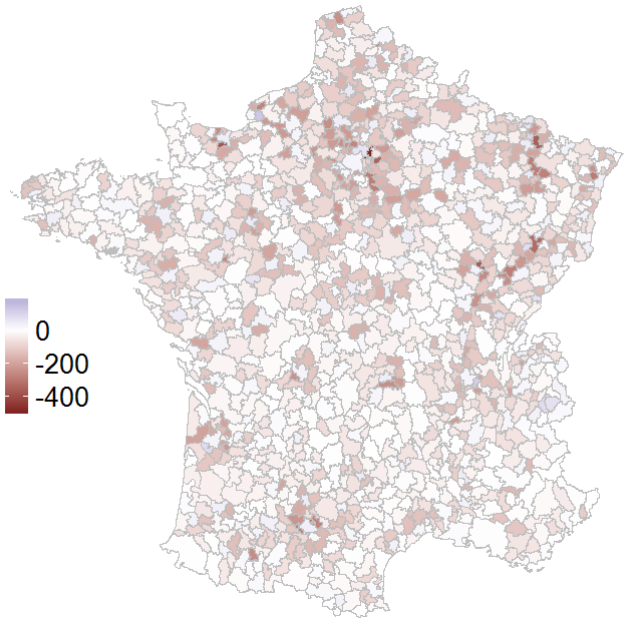
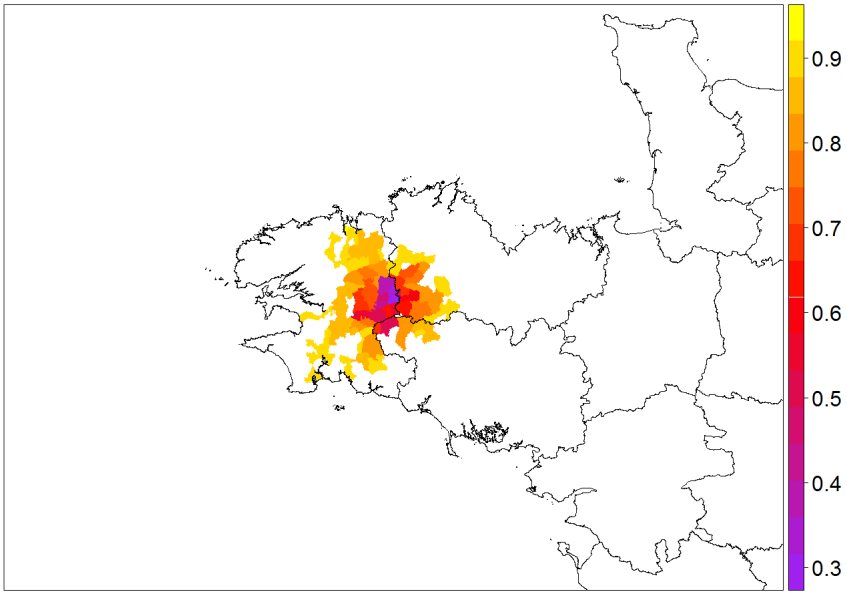


FIGURE 4.A.4 – Représentation du BEL de Carhaix-Plouguer (Algorithme ABSAPE)



Conclusion générale

Les débats ayant accompagné la redéfinition des régions en 2015 ont montré la difficulté de s'appuyer sur des critères objectifs pour agréger les territoires. La difficulté était accrue car des départements appartenant à une même région ne pouvaient être séparés. Même si cette exigence simplifiait les contraintes administratives liées à la mise en place de la réforme, elle mettait également en évidence la complexité d'agréger des blocs de départements [11]. Au regard des nouvelles compétences de certaines collectivités territoriales, cette question se révèle cruciale. En effet, si un périmètre est mal défini la décision politique qui peut impacter ce territoires n'est pas nécessairement du ressort des élus qui prennent les décisions. Ces questions relèvent du choix du découpage géographique pertinent : quels territoires sont à agréger et comment les agréger ?

La connaissance des facteurs de localisation permet de mieux objectiver la délimitation des périmètres territoriaux et d'offrir aux décideurs locaux des éléments objectifs d'appréciation.

L'opportunité d'améliorer la connaissance des territoires est indiscutable pour les décideurs publics comme pour les entreprises. Mais les entreprises peuvent, à la différence des collectivités territoriales, choisir leur localisation et en changer au fil du temps. Les firmes peuvent aussi se situer en plusieurs endroits pour mieux couvrir le territoire. Les firmes doivent également concilier le choix de la localisation qui permet de produire à bas coûts tout en garantissant un accès vers un marché final, ou plusieurs. Les déterminants utiles au choix de localisation peuvent dépendre du niveau géographique des données si elles sont agrégées.

Que ce soit pour des études empiriques ou des modélisations théoriques, l'intégration de l'espace et de la géographie se généralise dans les études. Cela peut être pour étudier l'organisation territoriale ou la concentration des activités, ou même dans le cadre de l'intégration des variables spatialisées pour des sujets d'ordres non géographiques. Tout au long de ce travail, nous nous sommes attachés à nous questionner sur les meilleures méthodes pour restituer l'espace et prendre en compte la géographie pour des problématiques économiques. Cela rend les études plus pertinentes empiriquement et théoriquement mais l'ajout de la dimension spatiale doit se faire avec des méthodes adaptées. La rigueur nécessaire à la prise en compte de l'espace peut être mise en

11. De nombreuses variantes de découpages se sont développées. Certaines étaient basées sur les équipements accessibles, d'autres sur des flux de mobilités domicile-travail, etc.

parallèle avec celle intégrant la dimension temporelle, même si cette dernière est plus ancienne (Thisse, 1997). L'amélioration technologique des outils à disposition permet de développer de nouvelles techniques pouvant s'affranchir d'une discrétisation de l'espace et de l'intégration de la notion explicite de distance. Les modèles de régression spatiale et les indices continus de concentration en sont des exemples. Plusieurs solutions sont mentionnées dans la littérature pour prendre en compte l'espace dans les études économiques, notamment la pondération des variables selon la contiguïté des zones. Une partie de cette thèse propose également des solutions et des recommandations pour mieux intégrer l'espace dans les études économiques.

Au cours de la thèse, nous avons mis en évidence des erreurs de mesure liées à l'utilisation d'un découpage géographique. Nous avons ainsi vu que les méthodes qui s'affranchissent de ces découpages, comme les indices continus de concentration, peuvent être sensibles à d'autres variables géographiques, notamment à la position des clusters sur le territoire.

Nous avons mené, au sein des trois premiers chapitres, une réflexion pour développer un outil opérationnel permettant de définir le périmètre géographique pertinent d'un territoire en limitant au maximum la sensibilité au découpage géographique.

Résumé des principaux résultats

Dans le premier chapitre, nous avons montré l'importance du choix de l'indice adéquat pour rendre compte de la concentration des activités dans l'espace. Aussi bien empiriquement qu'analytiquement, l'agrégation géographique entraîne systématiquement une hausse (indice de Herfindahl) ou une baisse (indice de Gini) de la concentration mesurée par les indices de première génération. Les indices de seconde génération, plus spécifiquement destinés à mesurer la concentration géographique des établissements, sont plus adaptés car ils prennent en compte l'agglomération globale de l'activité sur les différents territoires. Nous avons également constaté que les indices de seconde génération prennent en compte les interactions des logiques de localisation de différents secteurs. En faisant varier de manière systématique des frontières partitionnant l'espace sur une même répartition d'établissements, puis en faisant varier le nombre de régions, nous avons vu que l'indice de Maurel et Sédillot (1999) est l'indice discret le plus robuste pour restituer les valeurs attendues et celui qui semble être le moins sensible au MAUP malgré une variation plus importante que l'indice d'Ellison et Glaeser (1999) lorsque les découpages géographiques sont atypiques. Du fait de cette sensibilité au MAUP, le choix d'un indice selon le découpage géographique considéré n'est pas anodin lorsqu'il s'agit de comparer des secteurs ou de les prendre en compte dans des modèles de choix de localisation.

Le deuxième chapitre se focalise sur les mesures continues de concentration géographique des activités. Les apports de ces méthodes continues sont considérables pour les études impliquant économie et géographie, et plus particulièrement pour la localisation des activités. Ces méthodes permettent en effet d'en finir avec une vision discrète du territoire qui se résumait à un comptage d'observations dans chaque zone géographique.

Néanmoins la mise en application de ces méthodes se confrontent à plusieurs difficultés. Ces dernières peuvent être liées aux données, particulièrement lorsqu'il s'agit de réaliser l'arbitrage entre la précision géographique et le détails des observations. Les choix des différents paramètres des indices peuvent influencer les mesures de concentration, et comme pour les indices discrets, les mesures continues sont sensibles au nombre d'établissements du secteur. De nombreux travaux s'attachent à comparer la concentration des secteurs entre les différents pays. De ce point de vue, la catégorisation des établissements à partir d'une même classification sectorielle semble un prérequis. Ensuite, dans un contexte où l'évaluation de la "proximité" n'est pas appréhendée à partir du même choix de distance, la comparaison des territoires nécessite une clarification car même si les comparaisons qualitatives peuvent ne pas être affectées, les conclusions qualitatives peuvent être sujettes à caution. Enfin, la comparaison de secteurs n'ayant pas le même nombre d'établissements nous semble très délicate au regard des simulations effectuées.

Le troisième chapitre avait pour but de comparer différentes spécifications économétriques à un modèle de localisation contrôlé. Pour vérifier la conformité de ces spécifications, nous avons simulé un modèle de localisation à partir du modèle théorique de Weber (1929), afin de vérifier si les résultats estimés à partir de ces spécifications sont conformes à notre modèle initial. Notre approche nous a permis de mettre en lumière la forte corrélation des résultats des spécifications. Nous avons trouvé que la logique de localisation des secteurs d'inputs peut fortement influencer les résultats des estimations. Selon les logiques de localisation, une homogénéité ou une hétérogénéité des logiques de localisation peut améliorer la conformité des résultats au modèle de localisation. Le problème a de nouveau été mis en évidence car les résultats, dès lors que l'on utilise des variables géographiquement agrégées, sont sensibles à la manière dont le territoire est divisé.

La quatrième et dernière section est une proposition d'agrégation de territoires à partir d'un lieu de référence, afin d'établir le bassin économique local (BEL). Ce zonage, obtenu à partir des flux domicile-travail par commune, sert à identifier les spectateurs non locaux quand un événement est organisé. Au lieu de créer un nouveau zonage du territoire français, donc sensible au découpage géographique, nous considérons que chaque territoire peut appartenir à plusieurs BEL. Notre proposition ne permet pas une catégorisation stricte des territoires donc il n'est pas possible de

représenter ces BEL en une seule carte, mais il est possible d'établir une carte pour chaque BEL. Les deux processus itératifs développés mettent en évidence l'incapacité des découpages existants, plus particulièrement les découpages administratifs, à restituer la réalité des liens existants entre les territoires. Nous illustrons également notre méthode en considérant différents territoires et mettons en évidence que des territoires financent parfois des événements dans des communes qui ne font pas partie de leur BEL.

Limites et prolongement des travaux

Les travaux de cette thèse constituent également des étapes de recherches à approfondir. Les difficultés rencontrées et les questions soulevées permettent d'envisager des prolongements, ou des variantes, pour chaque chapitre.

Pour le chapitre 1 par exemple, nous avons limité le nombre d'indices à tester aux indices discrets utilisés dans la littérature. Nous pourrions élargir le calcul d'indices de concentration sur des localisations simulées par des indices d'autres domaines. Les indices d'entropie ou des indices de ségrégation sont des alternatives évoquées pour mesurer la concentration géographique des activités. Nous avons mené des tests pour les indices de Theil (1967) et de Duncan et Duncan (1955). Si les résultats ne sont pas meilleurs que ceux des indices déjà testés, ils apportent des informations relativement similaires à ces derniers. Il est donc possible qu'un indice appliqué à d'autres problématiques puisse être adapté à la mesure de la concentration géographique. Kubrak (2013) liste un nombre important d'indices de répartition et de spécialisation. Nous pourrions envisager de tester ces indices, ou des variantes, pour les adapter à la mesure de concentration géographique des établissements. Parmi ces tests, l'ajout d'une vérification analytique de l'agrégation sectorielle (MSUP) est la piste prioritaire envisagée. À partir de la base de données déjà utilisée, nous avons constaté que pour un même indice l'agrégation sectorielle et l'agrégation géographique ont empiriquement des effets inverses.

En parallèle, adopter une démarche plus empirique est possible. Avec les indices de Herfindahl, d'Ellison-Glaeser et de Maurel-Sédillot, il est possible d'identifier les territoires qui contribuent, à la baisse ou à la hausse, à la valeur finale de l'indice car les formules de ces indices correspondent à une somme de valeurs pour chaque territoire. La contribution de chaque territoire à la valeur finale est notamment montrée par Houdebine (1999). Outre une analyse spécifique d'un secteur, nous pourrions établir quel est le lien entre concentration et spécialisation des activités. Si nous avons strictement distingué les termes de concentration et de spécialisation, vérifier la corrélation, voire la causalité entre ces phénomènes est une piste de recherche prometteuse. Enfin, de la même manière que Briant *et al.* (2010), l'évaluation empirique de l'effet du MAUP sur d'autres pays que

la France constitue une piste possible pour un travail futur. Cela permettrait notamment de savoir si l'effet du MAUP est le même dans les différents pays de la nomenclature NUTS, nomenclature évoquée dans l'introduction générale.

Comme pour le chapitre 1, la méthodologie proposée dans le chapitre 2 peut être appliquée à d'autres indices continus. Marcon et Puech (2017) proposent plusieurs indices possibles à tester. Néanmoins, nous préférerions pouvoir appliquer les indices aux établissements d'un territoire, et vérifier sur données empiriques la sensibilité des conclusions obtenues. Plus spécifiquement, nous évoquons à la fois le MSUP en calculant des indices discrets et la sensibilité des indices continus au choix des localisations qui auraient pu être choisies. Il nous parait intéressant d'évaluer la sensibilité des indices continus au MSUP avec des données empiriques, et non simulées. Si les données sont accessibles à l'échelle communale, la réelle difficulté de cette recherche est liée au temps de calcul nécessaire pour effectuer toutes les combinaisons possibles de localisations. La limite à cette prolongation suggérée est due à des contraintes technologiques liées au temps de calcul induits.

De nombreuses variantes existent pour le chapitre 3. En effet, nous avons choisi un modèle de localisation à la Weber car il ne dépend d'aucun découpage géographique. Mais il est tout à fait envisageable d'intégrer d'autres facteurs de localisation qui peuvent dépendre de l'appartenance à une zone spécifique (comme un taux de taxe), ou d'intégrer une fonction de production avec une substituabilité des facteurs. Dans ce cas, une modélisation plus poussée devra être adossée au modèle de localisation. Calculer des distances euclidiennes est un choix qui peut aussi être remis en cause car il ne considère aucune spécificité géographique pouvant influencer le transport sur le territoire. Il est possible d'intégrer un réseau routier sur le territoire, créer une topographie pour le territoire, et même empêcher le passage sur certaines zones ou le rendre plus coûteux (fleuve, mer ou montagne).

Tous les prolongements de ce chapitre seraient d'autant plus pertinents si l'on pouvait adosser notre modèle à une application empirique. Dans ce cas, nous pourrions évaluer l'impact d'une politique publique sur les choix d'implantation des établissements (taxes favorables, nouvelles routes,...).

Nous avons occulté, dans ce chapitre, la notion de concentration géographique et les externalités possibles, qui était la notion pourtant centrale des précédents chapitres. Pour prendre en compte ce phénomène, il faudrait modifier le modèle de localisation et procéder à la localisation des établissements de manière séquentielle. C'est-à-dire qu'il faudrait implanter le premier établissement puis recalculer les profits possibles en chaque point, implanter un nouvel établissement, etc. Le risque

étant que la première localisation influence trop fortement les localisations suivantes. Compte tenu des contraintes techniques, l'ajout d'un facteur de concentration constituerait un travail à part entière. Il serait plus intéressant, dans ce cas, de reproduire les localisations existantes à une période antérieure, mesurer les déterminants de la localisation, et ensuite reproduire le modèle de localisation estimé mais en changeant les localisations de certains établissements (selon le sujet d'étude). Il serait ensuite possible d'évaluer l'effet de ces changements de localisations en comparant les localisations existantes et les localisations contrefactuelles obtenues suite à la modification de certaines localisations.

Concernant le chapitre 4, les développements futurs portent à la fois sur des variantes des algorithmes ainsi que sur l'utilisation des BEL créés. La création des algorithmes d'agrégation des territoires repose sur des hypothèses fortes ; en conséquence, de nombreuses variantes peuvent être testées. Les variantes envisagées sont liées à l'hypothèse de contigüité des zones agrégées, et donc à l'hypothèse que la zone créée est d'un seul tenant. Il est possible d'effectuer des variantes en prenant en compte non plus les zones adjacentes mais les zones en-dessous d'une distance seuil. En fait, les possibilités d'agréger les territoires sont similaires à la définition de voisinage pour la matrice de contigüité. Si l'on se base sur les conclusions de LeSage et Pace (2014), les variantes risquent néanmoins de donner des résultats similaires. L'objectif de la création des BEL est de permettre l'évaluation économique de l'organisation d'un événement à partir de la catégorisation des spectateurs en "locaux" et "non locaux". En conséquence, nous voulons utiliser ces BEL au cours d'une étude d'impact. Cependant, les données nécessaires pour appliquer notre méthode sont difficilement accessibles. Enfin, certains biens ou services ne peuvent avoir lieu qu'en présence de certains équipements (concerts,...). En conséquence, même si la motivation initiale conduit à des zones différentes des zones de chalandise, nous croyons qu'un rapprochement est possible avec d'autres domaines de recherche.

Liste des figures

Liste des tables

Bibliographie générale

INSEE Nord-Pas-de-Calais : Les déplacements domicile-travail entre les zones d'emploi du nord-pas-de-calais. *Dossier de Profils*, (102), 2011.

Lahsen Abdelmalki, Jean-Pierre Allegret, Florence Puech, Mustapha Sadni Jallab et Ahmed Silem : *Développements récents en économie et finances internationales : Mélanges en l'honneur du Professeur René Sandretto*. Armand Colin, 2012.

Karl Aiginger et Stephen W Davies : Industrial specialisation and geographic concentration : two sides of the same coin ? not for the european union. *Journal of Applied Economics*, 7(2):231–248, 2004.

José M Albert, Marta R Casanova et Vicente Orts : Spatial location patterns of spanish manufacturing firms. *Papers in Regional Science*, 91(1):107–136, 2012.

Christel Aliaga : Les zonages d'étude de l'insee : une histoire des zonages supracommunaux définis à des fins statistiques. 2015.

William Alonso : *Location and land use*. Harvard University Press Cambridge, MA, 1964.

Takeshi Amemiya : Qualitative response models : A survey. *Journal of economic literature*, 19 (4):1483–1536, 1981.

Josep-Maria Arauzo-Carod : Industrial location at a local level : comments on the territorial level of the analysis. *Tijdschrift voor economische en sociale geografie*, 99(2):193–208, 2008.

Josep-Maria Arauzo-Carod, Daniel Liviano-Solis et Miguel Manjón-Antolín : Empirical studies in industrial location : an assessment of their methods and results. *Journal of Regional Science*, 50(3):685–711, 2010.

Giuseppe Arbia : Modelling the geography of economic activities on a continuous space. *Papers in Regional Science*, 80(4):411–424, 2001.

Giuseppe Arbia et Giuseppe Espa : *Statistica economica territoriale*. Cedam, 1996.

Giuseppe Arbia, Giuseppe Espa et Danny Quah : A class of spatial econometric methods in the empirical analysis of clusters of firms in the space. *Empirical Economics*, 34(1):81–103, 2008.

Robert A Baade : The economic impact of mega-sporting events. *Handbook on the Economics of Sport*, pages 177–182, 2006.

Robert A Baade et Victor A Matheson : Going for the gold : The economics of the olympics. *Journal of Economic Perspectives*, 30(2):201–18, 2016.

Eric Barget et Alain Ferrand : Impact économique des événements sportifs sur le territoire : une méthode d'analyse basée sur les échanges entre les parties prenantes. *Management & Avenir*, (7):96–112, 2012.

Muriel Barlet, Anthony Briant et Laure Crusson : Concentration géographique dans l'industrie manufacturiere et dans les services en france : une approche par un indicateur en continu. *Documents de Travail de la DESE-Working Papers of the DESE*, 2008.

Muriel Barlet, Anthony Briant et Laure Crusson : Location patterns of service industries in france : A distance-based approach. *Regional Science and Urban Economics*, 43(2):338–351, 2013.

Muriel Barlet et Clémentine Collin : Localisation des professionnels de santé libéraux. *Comptes nationaux de la santé 2009*, 2010.

Robert J Barro et Xavier Sala-i Martin : Convergence. *Journal of political Economy*, 100 (2):223–251, 1992.

Kristian Behrens et Théophile Bougna : An anatomy of the geographical concentration of canadian manufacturing industries. *Regional Science and Urban Economics*, 51:47–69, 2015.

Moshe E Ben-Akiva, Steven R Lerman et Steven R Lerman : *Discrete choice analysis : theory and application to travel demand*, volume 9. MIT press, 1985.

J Besag : Contribution to the discussion on dr ripley's paper. *JR Stat. Soc.*, 39:193–195, 1977.

Stephen B Billings et Erik B Johnson : Measuring agglomeration : Which estimator should we use ? *Available at SSRN 2693098*, 2015.

Anthony Briant, Pierre-Philippe Combes et Miren Lafourcade : Dots to boxes : Do the size and shape of spatial units jeopardize economic geography estimations ? *Journal of Urban Economics*, 67(3):287–302, 2010.

Marius Brülhart et Rolf Traeger : An account of geographic concentration patterns in europe. *Regional Science and Urban Economics*, 35(6):597–624, 2005.

Maël-Luc Buron et Maëlle Fontaine : Confidentialité des données spatiales. *Manuel d'analyse spatiale*, pages 359–385, 2018.

Andrew J Cassey et Ben O Smith : Simulating confidence for the ellison–glaeser index. *Journal of Urban Economics*, 81:85–103, 2014.

Steven Coissard : Perspectives. la nouvelle économie géographique de paul krugman. *Revue d'Économie Régionale & Urbaine*, (1):111–125, 2007.

Pierre-Philippe Combes et Miren Lafourcade : Transport costs : measures, determinants, and regional policy implications for france. *Journal of economic geography*, 5(3):319–349, 2005.

Pierre-Philippe Combes et Henry G Overman : The spatial distribution of economic activities in the european union. *Handbook of regional and urban economics*, 4:2845–2909, 2004.

Jean-Philippe Croizean et Dany Vyt : Temps subjectif et temps mesuré : faut-il revoir la définition des zones de chalandise ? *Géographie, économie, société*, 17(2):201–224, 2015.

John L Crompton : Economic impact analysis of sports facilities and events : Eleven sources of misapplication. *Journal of sport management*, 9(1):14–35, 1995.

Matthieu Crozet et Miren Lafourcade : *La nouvelle économie géographique*. La Découverte, 2010.

Michael P Devereux, Rachel Griffith et Helen Simpson : The geographic distribution of production activity in the uk. *Regional Science and Urban Economics*, 34(5):533–564, 2004.

Magali Di Salvo, Monique Gadais et Geneviève Roche-Woillez : *L'estimation de la densité par la méthode du noyau : méthode et outils. Note méthodologique et technique*. CERTU, 2005.

Peter J Diggle et Amanda G Chetwynd : Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*, pages 1155–1163, 1991.

Otis Dudley Duncan et Beverly Duncan : A methodological analysis of segregation indexes. *American sociological review*, 20(2):210–217, 1955.

Gilles Duranton : La nouvelle économie géographique : agglomération et dispersion. *Économie & prévision*, 131(5):1–24, 1997.

Gilles Duranton et Henry G Overman : Testing for localization using micro-geographic data. *The Review of Economic Studies*, 72(4):1077–1106, 2005.

Glenn Ellison et Edward L Glaeser : Geographic concentration in us manufacturing industries : a dartboard approach. 1994.

Glenn Ellison et Edward L Glaeser : Geographic concentration in us manufacturing industries : A dartboard approach. *Journal of Political Economy*, 105(5):889–927, 1997.

Glenn Ellison et Edward L Glaeser : The geographic concentration of industry : does natural advantage explain agglomeration ? *The American Economic Review*, 89(2):311–316, 1999.

Glenn Ellison, Edward L Glaeser et William R Kerr : What causes industry agglomeration ? evidence from coagglomeration patterns. *The American Economic Review*, 100(3):1195–1213, 2010.

Pascal Eusebio, David Levy et Jean-Michel Floch : Partitionnement et analyse de graphes. *Manuel d'analyse spatiale*, pages 337–358, 2018.

Rodolphe Dos Santos Ferreira et Jacques-François Thisse : Horizontal and vertical differentiation : The launhardt model. *International Journal of Industrial Organization*, 14(4):485–506, 1996.

Jean-Michel Floch : Vivre en deçà de la frontière, travailler au-delà. *Insee Première*, 1337:1–4, 2011.

Richard Florida et Donald Smith : Toward the learning region. *Futures*, 27(5):527–536, 1995.

Ugo Fratesi : Issues in the measurement of localization. *Environment and Planning A*, 40(3):733–758, 2008.

Masahisa Fujita, Paul R Krugman et Anthony J Venables : *The spatial economy : cities, regions and international trade*, volume 213. Wiley Online Library, 1999.

Masahisa Fujita et Jacques-Francois Thisse : Spatial competition with a land market : Hotelling and von thunen unified. *The Review of Economic Studies*, 53(5):819–841, 1986.

Masahisa Fujita et Jacques-François Thisse : Economics of agglomeration : Cities. *Industrial Location, and Regional Growth, Cambridge*, 2002.

Robert C Geary : The contiguity ratio and statistical mapping. *The incorporated statistician*, 5 (3):115–146, 1954.

Steve Gibbons, Max Nathan et Henry Overman : Evaluating spatial policies. *Town Planning Review*, 85(4):427–432, 2014.

Corrado Gini : Variabilità e mutabilità. *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome : Libreria Eredi Virgilio Veschi*, 1, 1912.

Diego Giuliani, Giuseppe Arbia et Giuseppe Espa : Weighting ripley ?s k-function to account for the firm dimension in the analysis of spatial concentration. *International Regional Science Review*, 37(3):251–272, 2014.

Paulo Guimaraes, Octávio Figueiredo et Douglas Woodward : Accounting for neighboring effects in measures of spatial concentration. *Journal of Regional Science*, 51(4):678–693, 2011.

JI Haaland, HJ Kind, KH Midelfart-Knarvik et J Torstensson : What determines the economic geography of europe ? centre for economic policy research. Rapport technique, Discussion paper, 2072, 1999.

Richard Harris, John Moffat et Victoria Kravtsova : In search of "w". *Spatial Economic Analysis*, 6(3):249–270, 2011.

Fabrice Hatem : Attractivité : de quoi parlons-nous. *Pouvoirs locaux*, (61):39–40, 2004.

Jan Hauke et Tomasz Kossowski : Comparison of values of pearson's and spearman's correlation coefficient on the same sets of data. *Quaestiones Geographicae*, 30(2), 2011.

Keith Head, John Ries et Deborah Swenson : Agglomeration benefits and location choice : Evidence from japanese manufacturing investments in the united states. *Journal of international economics*, 38(3):223–247, 1995.

Vernon Henderson et Randy Becker : Political economy of city sizes and formation. *Journal of Urban Economics*, 48(3):453–484, 2000.

Orris C Herfindahl : *Concentration in the steel industry.* Thèse de doctorat, Columbia University., 1950.

Edgar M Hoover : Spatial price discrimination. *The Review of Economic Studies*, 4(3):182–191, 1937.

Harold Hotelling : Stability in competition. *Economic Journal*, 39:41–57, 1929.

Michel Houdebine : Concentration géographique des activités et spécialisation des départements français. *Economie et statistique*, 326(1):189–204, 1999.

INSEE : Guide du secret statistique. *Documentation INSEE*, 2010.

Walter Isard : The general theory of location and space-economy. *The Quarterly Journal of Economics*, 63(4):476–506, 1949.

Thomas Klier et Daniel McMillen : Plant location patterns in the european automobile supplier industry. *Growth and Change*, 46(4):558–573, 2015.

Thomas Klier et Daniel P McMillen : Evolving agglomeration in the us auto supplier industry. *Journal of Regional Science*, 48(1):245–267, 2008.

Reinhold Kosfeld, Hans-Friedrich Eckey et Jørgen Lauridsen : Spatial point pattern analysis and industry concentration. *The Annals of Regional Science*, 47(2):311–328, 2011.

Paul R Krugman : *Geography and trade*. MIT press, 1991a.

Paul R Krugman : The move toward free trade zones. *Economic Review*, 76(6):5, 1991b.

Paul R Krugman : *Development, geography, and economic theory*, volume 6. MIT press, 1997.

Claire Kubrak : Concentration et spécialisation des activités économiques : des outils pour analyser les tissus productifs locaux. *Document de travail Insee, E*, pages 213–1, 2013.

Miren Lafourcade et Giordano Mion : Concentration, agglomeration and the size of plants. *Regional Science and Urban Economics*, 37(1):46–68, 2007.

Wilhelm Launhardt : *Mathematische Begründung der Volkswirthschaftslehre*. W. Engelmann, 1885.

James P LeSage et R Kelley Pace : The biggest myth in spatial econometrics. *Econometrics*, 2 (4):217–249, 2014.

Dong Lin, Andrew Allan et Jianqiang Cui : The impacts of urban spatial structure and socio-economic factors on patterns of commuting : a review. *International Journal of Urban Sciences*, 19(2):238–255, 2015.

Eric Marcon et Florence Puech : Evaluating the geographic concentration of industries using distance-based methods. *Journal of Economic Geography*, 3(4):409–428, 2003.

Eric Marcon et Florence Puech : Measures of the geographic concentration of industries : improving distance-based methods. *Journal of Economic Geography*, 10(5):745–762, 2010.

Eric Marcon et Florence Puech : Mesures de la concentration spatiale en espace continu : théorie et applications. *Economie et statistique*, 474(1):105–131, 2014.

Eric Marcon et Florence Puech : A typology of distance-based measures of spatial concentration. *Regional Science and Urban Economics*, 62:56–67, 2017.

Alfred Marshall : Principles of political economy. *Maxmillan, New York*, 1890.

Bertil Matérn : Spatial variation : stochastic models and their application to some problems in forest surveys and other sampling investigations. *Meddelanden Fran Statens Skogsforskninginstitut*, 49(5), 1960.

Victor Matheson : Economic impact analysis. *Handbook on the Economics of Sport*, pages 137–142, 2006.

Françoise Maurel et Béatrice Sédillot : A measure of the geographic concentration in french manufacturing industries. *Regional Science and Urban Economics*, 29(5):575–604, 1999.

Eric Maurence : La mesure de l'impact économique d'un événement touristique. *Rapport d'étude, Paris, ministère de l'Économie, des Finances et de l'Industrie, DGCIS, Sous-direction de la Prospective, des Études économiques et de l'Évaluation*, 2012.

Thierry Mayer, Florian Mayneris et Loriane Py : The impact of urban enterprise zones on establishment location decisions and labor market outcomes : evidence from france. *Journal of Economic Geography*, 17(4):709–752, 2015.

Daniel McFadden : Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, pages 105–142, 1974.

Patrick AP Moran : Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

Leon N Moses : Location and the theory of production. *The Quarterly Journal of Economics*, 72 (2):259–272, 1958.

Kentaro Nakajima, Yukiko Umeno Saito et Iichiro Uesugi : Measuring economic localization : Evidence from japanese firm-level data. *Journal of the Japanese and International Economies*, 26(2):201–220, 2012.

Yann Nicolas : Les premiers principes de l'analyse d'impact économique local d'une activité culturelle. *Culture méthodes*, (1):1–8, 2007.

Sébastien Oliveau et Yoann Doignon : La diagonale se vide ? analyse spatiale exploratoire des décroissances démographiques en france métropolitaine depuis 50 ans. *Cybergeo : European Journal of Geography*, 2016.

Stan Openshaw : *The modifiable areal unit problem*, volume CATMOG 38. GeoBooks, Norwich, England, 1984.

Stan Openshaw et Peter J Taylor : A million or so correlation coefficients : three experiments on the modifiable areal unit problem. *Statistical applications in the spatial sciences*, 21:127–144, 1979.

Stan Openshaw et Peter J Taylor : The modifiable areal unit problem. *In* :quantitative geography : a british view.(eds n wrigley, r bennett) pp. 60–69, 1981.

Dominique Peeters et Jacky Perreur : L'approche weberienne de la localisation industrielle et ses extensions : un bilan. *L'Espace géographique*, pages 273–287, 1996.

James V Pinto : Launhardt and location theory : Rediscovery of a neglected book. *Journal of Regional science*, 17(1):17–29, 1977.

Pascal Pons et Matthieu Latapy : Computing communities in large networks using random walks. *J. Graph Algorithms Appl.*, 10(2):191–218, 2006.

Michael E Porter : The competitive advantage of notions. *Harvard business review*, 68(2):73–93, 1990.

Michael E Porter : Cluster and the new economics of competition. *Harvard College of Business*, 76(6), 1998.

Michael E Porter : Location, competition, and economic development : Local clusters in a global economy. *Economic development quarterly*, 14(1):15–34, 2000.

Florence Puech : *Concentration géographique des activités industrielles : Mesures et enjeux*. Thèse de doctorat, Paris 1, 2003.

Brian D Ripley : The second-order analysis of stationary point processes. *Journal of applied probability*, 13(2):255–266, 1976.

Brian D Ripley : Modelling spatial patterns. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 172–212, 1977.

Stuart S Rosenthal et William C Strange : The determinants of agglomeration. *Journal of urban economics*, 50(2):191–229, 2001.

Donald Smith et Richard Florida : Agglomeration and industrial location : an econometric analysis of japanese-affiliated manufacturing establishments in automotive-related industries. *Journal of Urban Economics*, 1994.

Catherine Sourd : L'attractivité économique des territoires. attirer des emplois, mais pas seulement. *Insee Première*, (1416), 2012.

Lasse Steiner, Bruno Frey et Simone Hotz : European capitals of culture and life satisfaction. *Urban studies*, 52(2):374–394, 2015.

Henri Theil : Economics and information theory. Rapport technique, 1967.

J-F Thisse, James E Ward et Richard E Wendell : Some properties of location problems with block and round norms. *Operations Research*, 32(6):1309–1327, 1984.

Jacques-François Thisse : L'oubli de l'espace dans la pensée économique. *Région et Développement*, 6:13–39, 1997.

Christine Thomas-Agnan et Florent Bonneu : Measuring and testing spatial mass concentration with micro-geographic data. *Spatial Economic Analysis*, 10(3):289–316, 2015a.

Christine Thomas-Agnan et Florent Bonneu : Measuring and testing spatial mass concentration with micro-geographic data. *Spatial Economic Analysis*, 10(3):289–316, 2015b.

Godfried T Toussaint : Pattern recognition and geometrical complexity. *In Proceedings Fifth International Conference on Pattern Recognition*, 1980.

Lukas Van Wyk, Melville Saayman, Riaan Rossouw et Andrea Saayman : Regional economic impacts of events : A comparison of methods. *South African Journal of Economic and Management Sciences*, 18(2):155–176, 2015.

JH Von Thünen : Der isolierte staat. *Beziehung auf Landwirtschaft und Nationalökonomie*, 1826.

Alfred Weber : *Ueber den standort der industrien*, volume 2. J.C.B. Mohr, Tübingen, 1909.

Alfred Weber : *Theory of the Location of Industries*. University of Chicago Press, 1929.

Douglas P Woodward : Locational determinants of japanese manufacturing start-ups in the united states. *Southern Economic Journal*, pages 690–708, 1992.

Yuri M Zhukov et Brandon M Stewart : Choosing your neighbors : Networks of diffusion in international relations. *International Studies Quarterly*, 57(2):271–287, 2013.

UNIVERSITE BRETAGNE ECONOMIE LOIRE ET GESTION

Le Mans Université

Titre : Localisation des activités économiques et sensibilité à l'échelle géographique considérée

Mots clés : agglomération, économie géographique, statistiques spatiales, économie spatiale, MAUP

Résumé : Nous nous proposons d'évaluer la sensibilité des mesures de concentration des activités économiques relativement au découpage géographique utilisé. Même si la prise en compte de la géographie dans les travaux économiques est devenue un sujet d'intérêt, des insuffisances dans les méthodes appliquées persistent. Dans le cadre de cette thèse nous mettons en évidence les limites de plusieurs outils couramment appliqués, et plus particulièrement la manière dont le MAUP (Modifiable Areal Unit Problem) les affecte. Dans le premier chapitre nous focalisons notre attention sur les indices de concentration géographique qui servent à appréhender les logiques de localisation des entreprises. Nous constatons que les indices couramment calculés présentent un biais (positif ou négatif selon l'indice) lorsque les données sont agrégées géographiquement. Nous vérifions cela de manière analytique et en simulant des localisations d'établissements, ce qui permet de comparer la valeur des indices calculés relativement aux paramètres de simulation de localisations.

Dans le deuxième chapitre, nous mettons en évidence que les indices qui ne discrétisent pas l'espace, et qui donc ne sont pas sensibles au MAUP, sont néanmoins sensibles à d'autres variables qui rendent difficiles la comparaison des résultats, par exemple lorsque le nombre d'établissements des secteurs ne sont pas égaux. Le troisième chapitre est consacré aux méthodes d'estimation des facteurs de localisation. L'apport de ce travail est la comparaison des résultats avec un modèle de localisation que nous simulons selon plusieurs paramètres (logique de localisation, importance des facteurs, etc). Nous trouvons que la variété des logiques de localisations altère la qualité des estimations empiriques. Compte tenu du travail effectué jusqu'alors, le dernier chapitre propose un outil pour déterminer de manière objective la zone d'activités économiques pertinentes, outil qui n'est pas sensible au MAUP. Ce travail utilise les flux domicile-travail pour identifier les périmètres géographiques en lien avec un lieu de référence choisi. Cet outil opérationnel permet de segmenter les visiteurs entre « locaux » et « non locaux » dans le cadre des études d'impact économique.

Title : Location of economic activities and sensitivity to the geographic scale considered

Keywords : agglomeration, geographical economy, spatial statistics, spatial economy, MAUP

Abstract : We propose to evaluate the sensitivity of the concentration activities measures to the geographical scale used. Even if the consideration of geography in economic work has become a subject of interest, deficiencies in applied methods persist. In this thesis, we highlight the limits of several commonly applied tools, and more particularly the way in which the Modifiable Areal Unit Problem (MAUP) affects them. In the first chapter we focus our attention on the indices of geographic concentration that serve to apprehend the location strategy of firms. We find that the commonly calculated indices have a bias (positive or negative depending on the index) when the data are geographically aggregated. We test this analytically and by simulating location of establishments, which allows us to compare the value of the indices calculated with respect to the location simulation parameters.

In the second chapter, we highlight that indices that do not discretize space, and therefore are not sensitive to MAUP, are nevertheless sensitive to other variables that make it difficult to compare results, for example when the size of each sector are not equal. The third chapter is devoted to methods for estimating location factors. The contribution of this work is the comparison of the results with a localization model that we simulate according to several parameters (location strategy, importance of the factors, etc). We find that the variety of location strategy alters the quality of empirical estimates. In view of the work done so far, the last chapter proposes a tool to objectively determine the area of relevant economic activities, a tool that is not sensitive to MAUP. This work uses home-work flows to identify geographical perimeters in relation to a reference site. This operational tool is used to segment visitors from "local" or "non-local" in economic impact studies.