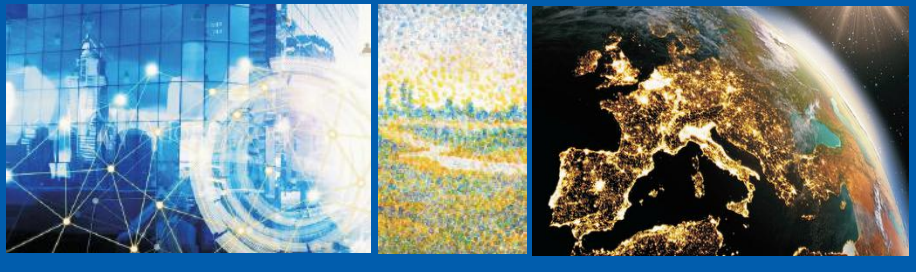
Insee Méthodes



N° 131
Octobre 2018

Manuel d'analyse spatiale

Théorie et mise en oeuvre pratique avec R

Insee - Eurostat
Sous la direction de Vincent LOONIS
Coordonné par Marie-Pierre de BELLEFON

Insee
Mesurer pour comprendre

eurostat

EUROPEAN FORUM FOR
GEOGRAPHY AND STATISTICS

MANUEL D'ANALYSE SPATIALE

Théorie et mise en œuvre pratique avec R

Ce projet a été en partie financé par le programme statistique européen 2013-2017 dans le cadre de l'action ESS "Intégration de l'information statistique et géospatiale" par la subvention numéro 08143.2015.001-2015.714

Direction	Vincent Loonis
Coordination	Marie-Pierre de Bellefon
Mise en cohérence éditoriale	Vianney Costemalle, Maëlle Fontaine
Contribution	*Insee :* Pascal Ardilly, Sophie Audric, Marie-Pierre de Bellefon, Maël-Luc Buron, Eric Durieux, Pascal Eusébio, Cyril Favre-Martinoz, Jean-Michel Floch, Maëlle Fontaine, Laure Genebes, Ronan Le Gleut, Raphaël Lardeux-Schutz, David Lévy, Vincent Loonis, Ronan Le Saout, Thomas Merly-Alpa, Auriane Renaud, François Sémécurbe *GAINS (TEPP) et Crest Le Mans Université* Salima Bouayad-Agha *AgroParisTech, UMR EcoFoG* Éric Marcon *RITM, Univ. Paris-Sud, Université Paris-Saclay et Crest* Florence Puech *CESAER, AgroSup Dijon, INRA, Université de Bourgogne Franche- Comté* Julie Le Gallo, Lionel Védrine *ENSAI* Paul Bouche, Wencan Zhu
Directeur de la publication	Jean-Luc Tavernier
Editeur	Institut national de la statistique et des études économiques 88, avenue Verdier - CS 700058 92541 Montrouge Cedex *www.insee.fr*

Table des matières

Éditorial EFGS - Eurostat

Mariana Kotzeva - *Directrice Générale d'Eurostat*
Janusz Dygaszewicz - *Président du Forum Européen pour la Géographie et la Statistique (EFGS), Institut de Statistique de Pologne*

Ces dernières années, diverses initiatives internationales et nationales ont été lancées pour jeter des ponts entre le monde des informations géospatiales et celui des observations statistiques.

Le 27 juillet 2011, le Conseil économique et social des Nations Unies (ECOSOC) a pris acte du besoin de promouvoir la coopération internationale dans le domaine des informations géospatiales et a décidé, à cet égard, d'établir le Comité d'experts sur la gestion de l'information géospatiale à l'échelle mondiale (UN-GGIM). Lors de sa troisième session, tenue au Royaume-Uni en juillet 2013, le Comité d'experts a adopté la décision 3/107 (voir document E/C.20/2013/17) qui "a reconnu qu'il était essentiel d'intégrer les informations géospatiales avec les statistiques et les données socioéconomiques et d'élaborer un cadre statistico-spatial".

Le cadre statistico-spatial mondial (CSSM) actuellement en cours de développement proposera une méthode commune, intégrée et interopérable, pour la prise en compte des informations géospatiales dans les statistiques et la gestion de ces informations géospatiales à toutes les étapes de la production statistique. Il fait le lien entre informations spatiales, qui décrivent notre environnement anthropique, naturel et physique, et informations statistiques, qui décrivent leurs attributs socio-économiques et environnementaux. Ce cadre s'est déjà révélé utile tant pour le Programme de développement durable à l'horizon 2030 que pour les recensements de population du cycle de 2020.

Eurostat soutient pleinement les initiatives mondiales et leur transposition à l'échelle européenne. La mise en œuvre européenne de la stratégie mondiale repose sur des lignes directrices méthodologiques élaborées dans le cadre de la série de projets Geostat. La série de projets Geostat a démarré avec l'ambition très concrète de cartographier le recensement de 2011 sur un carroyage de population à l'échelle de l'UE (Geostat 1). Le niveau d'ambition et le champ d'application ont été progressivement élargis par le développement, pour la statistique, d'un cadre de référence géospatial normalisé basé sur des points, fondé sur des registres d'adresses, d'immeubles et/ou de logements géocodés (Geostat 2) ou encore par la mise au point et le test du CSSM dans le contexte européen (Geostat 3).

L'organisation de ces projets a bénéficié des échanges d'idées fructueux alimentés par les conférences annuelles de l'*European Forum for Geography and Statistics* (EFGS, forum européen pour la géographie et la statistique) financé par Eurostat. Le principal objectif de l'EFGS est de promouvoir l'intégration des données statistiques et spatiales ainsi que l'utilisation des observations géospatiales dans les processus décisionnels. L'EFGS est un organisme bénévole qui réunit des spécialistes des données géospatiales mais également des statisticiens, des chercheurs travaillant dans les instituts nationaux de statistique (INS) et des experts des instituts nationaux de géographie. Des pays extérieurs à l'Europe participent également à l'EFGS et le Forum vise par conséquent à établir une plateforme mondiale et une solide coopération avec l'UN-GGIM aux niveaux régional et mondial. En 2015, l'EFGS a reçu le statut officiel d'Organisation observatrice auprès de l'UN-GGIM : Europe. Comme indiqué précédemment, l'EFGS agit également en tant que groupe de référence pour Eurostat et en tant que groupe de travail dans le cadre des projets Geostat.

Eurostat et l'EFGS conviennent que l'intégration des informations statistiques et spatiales doit reposer sur de solides conseils de méthodologie afin de garantir la qualité et la comparabilité des

statistiques géospatiales. Pour cette raison, les deux organisations saluent chaudement l'initiative de l'Insee qui vise la rédaction d'un manuel d'analyse spatiale, qui part d'un système d'information statistique dont tous les unités sont géolocalisées. Elles approuvent sans réserve les objectifs du manuel, qui sont de promouvoir, développer et consolider l'utilisation de méthodes statistiques spécifiques à disposition des INS, uniquement dans le cadre d'un tel système d'information. Ces méthodes, qui vont de l'évaluation de l'autocorrélation spatiale à la conception d'un échantillon spatialement équilibré en passant par la gestion de la confidentialité en contexte spatial, cadrent parfaitement avec la finalité des activités de n'importe quel INS. Un tel outil, qui se concentre sur des exemples pratiques et leur application dans le monde de la recherche, rendra sans aucun doute le processus de production statistique, la publication ou l'analyse des résultats statistiques plus efficaces.

Nous sommes convaincus que ce manuel sera très utile pour les spécialistes des instituts de statistique du monde entier et des instituts nationaux de géographie qui souhaitent en savoir plus sur l'utilisation des informations spatiales en statistique. Le manuel sera également précieux pour les producteurs et utilisateurs de données statistiques, qu'ils soient spécialistes ou débutants, en ce qu'il leur permettra d'identifier les opportunités d'intégration des données mais aussi de comprendre les défis méthodologiques que peut poser l'intégration de ces deux types de données, parfois très différentes. Nous espérons qu'il incitera beaucoup de spécialistes à utiliser les informations spatiales pour la production statistique, et à faire entrer les statistiques géospatiales dans les processus décisionnel et d'analyse.

Éditorial Insee - guide de lecture

Vincent Loonis - *Responsable de la Division des Méthodes et Référentiels Géographiques (DMRG), Insee*
Marie-Pierre de Bellefon - *Responsable de la Section Méthodes d'Analyse Spatiale à la DMRG*

Pourquoi un nouveau manuel d'analyse spatiale ?

Cressie fut un des premiers à publier un "manuel de statistiques spatiales" (CRESSIE 1993a). Clair et détaillé, son ouvrage permet d'approfondir la théorie des statistiques spatiales. Il n'inclut cependant aucun guide d'utilisation pratique de ces méthodes. Depuis cette publication, les avancées théoriques et informatiques sont allées de pair avec l'accroissement de l'offre de données géolocalisées. De nombreux spécialistes ont à leur tour rédigé des manuels et autres guides de statistique spatiale : des très théoriquesPACE et al. 2009, GELFAND et al. 2010 ou ANSELIN 2013 aux guides d'utilisation du logiciel R : BIVAND et al. 2008, BRUNSDON et al. 2015 en passant par des ouvrages mêlant théorie et pratique comme HAINING 2003, SCHABENBERGER et al. 2004 ou FISCHER et al. 2009. Parmi les ouvrages francophones, ZANINETTI 2005 décrit la théorie de la statistique spatiale, tandis que CALOZ et al. 2011 s'intéressent à la géostatistique. Au sein même de l'Insee, Jean-Michel Floch a présenté en 2013 l'apport de la statistique spatiale pour l'étude des disparités socio-économiques (FLOCH 2013) et en 2015 ses réflexions sur la statistique spatiale en général.

L'objectif du manuel d'analyse spatiale est de répondre aux questions concrètes des chargés d'étude des instituts statistiques : que faire avec ces nouvelles sources de données géolocalisées ? Dans quels cas doit-on prendre en compte leur dimension spatiale ? Comment appliquer les méthodes de statistique et d'économétrie spatiale ? Contrairement aux manuels existants, la pédagogie est pensée spécifiquement en fonction des enjeux propres aux instituts statistiques : les exemples d'application utilisent des données collectées par la statistique publique et l'accent est mis sur la pratique et l'importance du choix des paramètres. Les fondements théoriques sont suffisamment approfondis pour permettre de comprendre les subtilités dans la mise en œuvre pratique des méthodes, tout en renvoyant aux ouvrages spécialisés les lecteurs désireux de connaître les extensions d'un niveau technique plus élevé. La majorité des chapitres présente des méthodes bien documentées et fréquemment utilisées, mais quelques-uns s'appuient sur des travaux innovants diffusés récemment. Parmi les thèmes abordés, le manuel Insee-Eurostat s'intéresse aux questions de sondage et de respect de la confidentialité ; autant de points importants pour les INS et très peu approfondis dans les ouvrages existants. Quelques chapitres ouvrent sur des notions peu utilisées actuellement à l'Insee comme la géostatistique.

Le panel d'auteurs mêle statisticiens de différents départements de l'Insee (Département de la Méthodologie Statistique, Département de l'Action Régionale, Département des Études et Synthèses Économiques) et professeurs universitaires (universités du Mans, Paris-Sud, Guyane, Agrosup et Inra Dijon). La rédaction du manuel a ainsi été l'occasion de favoriser les échanges entre le milieu de la statistique publique et le milieu académique.

Le plan du manuel

En 2008, le prix Nobel d'économie fut remis à Paul Krugman : le père de la nouvelle économie géographique. Cette récompense marque l'importance croissante de la prise en compte des

phénomènes spatiaux. Krugman décrit l'économie géographique comme "la branche des sciences économiques qui s'intéresse au lieu où les choses se produisent et aux relations entre elles" (KRUGMAN 1991). Cette citation illustre la démarche propre à toute étude d'analyse spatiale, quel qu'en soit le domaine d'application. L'analyste commence par décrire les lieux des observations, puis il mesure l'importance des interactions spatiales afin de pouvoir prendre en compte ces interactions grâce à un modèle pertinent. Ces trois étapes correspondent aux trois premières parties du manuel : Partie 1 : *Décrire les données géolocalisées* ; Partie 2 : *Mesurer l'importance des effets spatiaux* ; Partie 3 : *Prendre en compte les effets spatiaux*.

Le lieu est référencé dans un système d'information géographique grâce à ses coordonnées. Une des caractéristiques de l'analyse spatiale est donc que le support de l'observation, défini comme l'ensemble des coordonnées spatiales des objets à traiter, contient des informations potentiellement riches pour l'analyse. Pour les exploiter, le chargé d'étude commence le plus souvent par regrouper les données en fonction de leur proximité géographique. Il s'agit de la première étape avant d'explorer les caractéristiques de la localisation des données et de décrire l'évolution des variables dans l'espace. Ce regroupement est aussi un paramètre clé pour assurer le respect de la confidentialité des données diffusées par les instituts de statistique publique. Le premier chapitre du manuel : *Analyse spatiale descriptive* présente la façon dont on peut prendre en main les données avec le logiciel R et réaliser de premières cartes. Des notions de sémiologie cartographique sont également introduites. La deuxième étape d'une analyse spatiale est la définition du voisinage d'un objet. La définition du voisinage est indispensable pour mesurer la force des relations spatiales entre les objets, c'est-à-dire la façon dont les voisins s'influencent les uns les autres. L'enjeu du deuxième chapitre du manuel : *Codifier la structure de voisinage*, est de réussir à définir des relations de voisinage cohérentes avec les véritables interactions spatiales entre les objets. Ce chapitre présente plusieurs notions de voisinage, fondées sur la contiguïté ou sur les distances entre observations. La question du poids accordé à chaque voisin est aussi abordée.

Les données géolocalisées peuvent être réparties en trois catégories : données surfaciques, données ponctuelles et données continues. La différence fondamentale entre ces données n'est pas la taille de l'unité géographique considérée mais le processus générateur des données. Pour des données surfaciques, la localisation des observations est considérée comme fixe : c'est la valeur des observations qui suit un processus aléatoire. Par exemple, le PIB de chaque région est une donnée spatiale surfacique. Plus les valeurs des observations sont influencées par les valeurs des observations qui leur sont géographiquement proches, plus l'autocorrélation spatiale est élevée. Les indices d'autocorrélation spatiale permettent de mesurer la force des interactions spatiales entre les observations. Les versions globales et locales des indices d'autocorrélation spatiale sont présentées dans le chapitre 3 : *Indices d'autocorrélation spatiale*. Pour des données ponctuelles, la localisation des observations est la variable aléatoire. Il peut s'agir par exemple de la localisation des commerces au sein d'une ville. La force des interactions spatiales se mesure donc à l'aune de l'écart entre la distribution dans l'espace des observations et une distribution spatiale complètement aléatoire.Le chapitre 4 : *Les configurations de points* donne les méthodes et les outils permettant notamment de mettre en évidence les éventuelles attractions ou répulsions entre les différents types de points et la façon dont on évalue la significativité des résultats obtenus. Enfin, les données continues se caractérisent par le fait qu'il existe une valeur pour la variable d'intérêt en tout point du territoire étudié. En revanche ces données sont mesurées uniquement en un nombre discret de points. Il s'agit, par exemple, de la composition chimique du sol utile à l'industrie minière. Le chapitre 5 : *Géostatistique* présente les concepts fondamentaux permettant d'étudier les données continues : semi-variogramme, interpolation des données par les méthodes de krigeage,...

Les troisièmes et quatrièmes parties du manuel se concentrent sur l'étude des données surfaciques, auxquelles on a le plus souvent affaire dans les instituts de statistique publique. Parmi

les phénomènes spatiaux qui affectent les données surfaciques, on peut distinguer dépendance spatiale et hétérogénéité spatiale. La dépendance spatiale désigne une situation où la valeur d'une observation est liée aux valeurs des observations voisines (soit elles s'influencent mutuellement, soit elles sont toutes les deux soumises à un même phénomène inobservé). L'économétrie spatiale modélise cette dépendance spatiale. Plusieurs formes d'interactions existent, relatives à la variable à expliquer, aux variables explicatives ou aux variables inobservées. De nombreux modèles se retrouvent donc en concurrence, à partir d'une même définition préalable des relations de voisinage. Le chapitre 6 : *Économétrie spatiale, modèles courants* détaille la méthodologie pas à pas de choix de modèle (estimation et tests), ainsi que les précautions à prendre dans l'interprétation des résultats. La façon dont les modèles d'économétrie spatiale peuvent être appliqués à l'étude des données de panel est présentée dans le chapitre 7 : *Économétrie spatiale sur données de panel*.

L'hétérogénéité spatiale désigne le fait que l'influence des variables explicatives sur la variable dépendante varie avec la localisation des observations. La régression géographiquement pondérée ou le lissage spatial prennent en compte ce phénomène. Indépendamment d'un modèle de régression, le *lissage spatial* (chapitre 8) filtre l'information pour révéler les structures spatiales sous-jacentes. La *régression géographiquement pondérée* (chapitre 9) répond plus précisément au constat qu'un modèle de régression estimé sur l'ensemble d'un territoire d'intérêt peut ne pas appréhender de façon adéquate les variations locales. La régression géographiquement pondérée permet, notamment à l'aide de représentations cartographiques associées, de repérer où les coefficients locaux s'écartent le plus des coefficients globaux, et de construire des tests permettant d'apprécier si et comment le phénomène est non stationnaire.

Qu'elles soient destinées à prendre en compte la dépendance ou l'hétérogénéité spatiale, les méthodes d'analyse spatiale ont été développées à partir de données exhaustives. Elles peuvent cependant enrichir l'éventail des techniques liées aux sondages. Ces techniques sont particulièrement importantes pour les instituts de statistique publique, dont les données sont souvent obtenues grâce à une enquête. En amont, la constitution des entités à sélectionner aux premiers degrés d'un plan de sondage et la sélection de l'échantillon peuvent être améliorées grâce aux techniques d'échantillonnage spatial présentées dans le chapitre 10. En aval, le chapitre 11 : *Économétrie spatiale sur données d'enquête* présente les écueils liés à l'estimation d'un modèle d'économétrie spatiale sur données échantillonnées et évalue les potentielles corrections proposées par la littérature empirique. Le chapitre 12 : *Estimation sur petits domaines et corrélation spatiale* présente les méthodes petits domaines et la façon dont la prise en compte de la corrélation spatiale peut améliorer les estimations.

La quatrième partie du manuel : *Prolongements* introduit deux chapitres qui utilisent directement la dimension spatiale des données, tout en s'éloignant du traitement classique de la dépendance ou l'hétérogénéité spatiale. L'analyse des réseaux permet de prendre en compte l'ensemble des flux entre les territoires pour déterminer les relations privilégiées. Les techniques *d'analyse et de partitionnement de graphes* sont présentées dans le chapitre 13. La profusion de données géocodées va de pair avec un risque de divulgation élevé, puisque le nombre de variables nécessaires pour identifier une personne de manière unique diminue considérablement lorsque l'individu auteur de l'intrusion connaît la position géographique précise. Ce sujet est crucial pour les instituts de statistiques qui sont soumis à de fortes demandes de diffusion de données sensibles à des niveaux géographiques toujours plus fins. Le chapitre 14 : *Confidentialité des données spatiales* vise à proposer des suggestions pour évaluer et gérer le risque de divulgation, tout en préservant les corrélations spatiales.

La lecture des trois premiers chapitres est recommandée pour faciliter la compréhension de l'ensemble du manuel. Le préambule de chaque chapitre précise ensuite les chapitres particuliers dont la lecture préalable est nécessaire à la bonne compréhension du chapitre. Le corps du texte

présente la théorie fondamentale et les exemples d'application pratique. Les encadrés sont des extensions plus techniques dont la lecture n'est pas impérative pour comprendre l'essentiel de la méthode.

Références - Editorial Insee

ANSELIN, Luc (2013). *Spatial econometrics : methods and models*. T. 4. Springer Science & Business Media.

BIVAND, Roger S., Edzer PEBESMA et Virgilio GOMEZ-RUBIO (2008). *Applied spatial data analysis with R*. Springer.

BRUNSDON, Chris et Lex COMBER (2015). *An Introduction to R for Spatial Analysis Et Mapping*. Sage London.

CALOZ, Régis et Claude COLLET (2011). *Analyse spatiale de l'information géographique*. PPUR Presses polytechniques.

CRESSIE, Noel (1993a). *Statistics for spatial data*. John Wiley & Sons.

FISCHER, Manfred M et Arthur GETIS (2009). *Handbook of applied spatial analysis : software tools, methods and applications*. Springer Science & Business Media.

FLOCH, Jean-Michel (2013). « Détection des disparités socio-économiques, l'apport de la statistique spatiale ».

GELFAND, Alan E et al. (2010). *Handbook of spatial statistics*. CRC press.

HAINING, Robert P (2003). *Spatial data analysis : theory and practice*. Cambridge University Press.

KRUGMAN, Paul R (1991). *Geography and trade*. MIT press.

PACE, R Kelley et JP LESAGE (2009). « Introduction to spatial econometrics ». *Boca Raton, FL : Chapman &Hall/CRC*.

SCHABENBERGER, Oliver et Carol A GOTWAY (2004). *Statistical methods for spatial data analysis*. CRC press.

ZANINETTI, Jean-Marc (2005). *Statistique spatiale : méthodes et applications géomatiques*. Hermès science publications.

Auteurs et relecteurs

Analyse spatiale descriptive

Auteurs : Sophie Audric, Marie-Pierre de Bellefon, Éric Durieux
Relecteurs : Maëlle Fontaine, François Sémécurbe

Codifier la structure de voisinage

Auteurs : Marie-Pierre de Bellefon, Ronan Le Gleut, Vincent Loonis
Relecteurs : Salima Bouayad-Agha, Ali Hachid

Indices d'autocorrélation spatiale

Auteurs : Marie-Pierre de Bellefon, Salima Bouayad-Agha
Relecteurs : Olivier Sautory, Lionel Védrine

Les configurations de points

Auteurs : Jean-Michel Floch, Éric Marcon, Florence Puech
Relecteurs : Salima Bouayad-Agha, Gabriel Lang

Géostatistique

Auteur : Jean-Michel Floch
Relecteurs : Marie-Pierre de Bellefon, Thomas Romary

Économétrie spatiale : modèles courants

Auteurs : Jean-Michel Floch, Ronan Le Saout
Relecteurs : Salima Bouayad Agha, Pauline Givord, Julie Le Gallo, Olivier Sautory

Économétrie spatiale sur données de panel

Auteurs : Salima Bouayad-Agha, Julie Le Gallo, Lionel Védrine
Relecteurs : Sébastien Faivre, Alain Pirotte

Lissage spatial

Auteurs : Laure Genebes, Auriane Renaud, François Sémécurbe
Relecteurs : Valérie Darriau, Jean-Michel Floch

Régression géographiquement pondérée

Auteurs : Marie-Pierre de Bellefon, Jean-Michel Floch
Relecteurs : Maëlle Fontaine, François Sémécurbe

Échantillonnage spatial

Auteurs : Cyril Favre-Martinoz, Maëlle Fontaine, Ronan Le Gleut, Vincent Loonis
Relecteurs : Éric Lesage, Patrick Sillard

Économétrie spatiale sur données d'enquête

Auteurs : Raphaël Lardeux-Schutz, Thomas Merly-Alpa
Relecteur : Ronan Le Saout

Estimation sur petits domaines et corrélation spatiale

Auteurs : Pascal Ardilly, Paul Bouche, Wencan Zhu
Relecteurs : Jean-François Beaumont, Olivier Sautory

Partitionnement et analyse de graphes

Auteurs : Pascal Eusébio, Jean-Michel Floch, David Lévy
Relecteurs : Laurent Beauguitte, Benjamin Sakarovitch

Confidentialité des données spatiales

Auteurs : Maël-Luc Buron, Maëlle Fontaine
Relecteurs : Maxime Bergeat, Heïdi Koumarianos

Remerciements

Nous remercions le comité de pilotage du manuel pour nous avoir aidés à définir les orientations à donner à ce manuel, et pour leur travail de relecture de l'ensemble du manuel : Salima Bouayad-Agha, Vianney Costemalle, Valérie Darriau, Marie-Pierre de Bellefon, Gaël de Peretti, Sébastien Faivre, Jean-Michel Floch, Maëlle Fontaine, Pauline Givord, Vincent Loonis, Olivier Sautory, François Sémécurbe, Patrick Sillard.

Nous remercions également Hicham Abbas, Jérôme Accardo, Kathleen Aubert, Maël-Luc Buron, Julie Djiriguian, Pascale Rouaud et Sonia Oujia pour leur aide dans la relecture de la version finale du manuel.

Enfin, nous remercions Brigitte Rols pour la conception de la couverture.

Partie 1 : décrire les données géolocalisées

1. Analyse spatiale descriptive

SOPHIE AUDRIC, MARIE-PIERRE DE BELLEFON, ERIC DURIEUX
Insee

Résumé

L'objectif de l'analyse spatiale est de comprendre et d'explorer l'intrication entre le positionnement spatial des objets et des phénomènes, et leurs caractéristiques. La littérature distingue traditionnellement trois types de données spatiales : données ponctuelles, données continues et données surfaciques. À chaque type de données correspondent des méthodes d'analyse spécifiques. Cependant, quelle que soit la nature des données spatiales, la première étape est de les prendre en main et de les agréger à une échelle géographique adaptée au processus spatial sous-jacent. Cartographier les données permet de synthétiser une information, de la rendre accessible à un public élargi et de réfléchir aux outils statistiques adaptés à la poursuite de l'étude. Cette première analyse descriptive peut également, dans la démarche d'une étude, être l'occasion de constater des problèmes particuliers dans les données (collecte, données manquantes, valeurs aberrantes,etc.) ou d'invalider certaines hypothèses nécessaires au développement de méthodes économétriques. Nous introduisons dans ce chapitre les notions de sémiologie cartographique utiles pour réaliser une carte de qualité. [1]

1. Ces éléments de sémiologie sont extraits d'un ouvrage de l'Insee "Guide de sémiologie cartographique" publié en 2017 et auquel ont contribué un grand nombre de personnes que nous remercions.

Ce chapitre décrit la prise en main des données spatiales avec le logiciel R et la production de premières cartes descriptives. Des études réalisées dans divers instituts de statistique européens illustrent ces notions.

1.1 Différents types de données spatiales

Une donnée spatiale est une observation dont on connaît non seulement la valeur, mais aussi la localisation. Le support des observations, défini comme l'ensemble des coordonnées spatiales des objets à traiter, constitue une information potentiellement riche pour l'analyse.
Certaines propriétés des données spatiales contredisent les hypothèses nécessaires à l'utilisation des méthodes statistiques habituelles. Ainsi, l'hypothèse d'indépendance des observations, requise dans la plupart des modèles économétriques, n'est pas vérifiée en présence de *dépendance spatiale* : lorsque la valeur de l'observation i influence la valeur de l'observation j voisine. Les données spatiales peuvent aussi se caractériser par de l'*hétérogénéité spatiale* : l'influence des variables explicatives sur la variable dépendante dépend de la localisation dans l'espace ; une variable peut-être influente sur une autre dans un voisinage donné, mais ne pas l'être dans un autre endroit. Pour analyser les données spatiales, de nombreuses méthodes spécifiques ont donc été développées.
Les méthodes et leurs objectifs dépendent de la nature des données spatiales. D'après la classification proposée par CRESSIE 1993b, on distingue trois types de données spatiales :

— données ponctuelles ;
— données continues ;
— données surfaciques.

La différence fondamentale entre ces données n'est pas la taille de l'unité géographique considérée mais le processus générateur des données.

1.1.1 Données ponctuelles

Les données spatiales ponctuelles se caractérisent par la **distribution dans l'espace** des observations. Le processus générateur des données génère les coordonnées géographiques associées à l'apparition d'une observation. On n'étudie pas de valeur associée à l'observation ; seule compte la localisation. Il s'agit, par exemple, du lieu d'apparition d'une maladie lors d'une épidémie, ou de la répartition dans l'espace de certaines espèces d'arbres. L'analyse spatiale des données ponctuelles a pour objectif de **quantifier l'écart entre la distribution spatiale des observations et une distribution complètement aléatoire dans l'espace**. Si les données sont plus regroupées que si elles étaient distribuées aléatoirement sur le territoire, on peut identifier des clusters et mesurer leur significativité.


Les principales méthodes permettant d'analyser les données ponctuelles sont décrites dans le chapitre 4 : "Configurations de points".

■ **Exemple 1.1 — Détection de clusters.** FOTHERINGHAM *et al. 1996* cherchent à détecter la présence de clusters significatifs de maisons inconfortables. Ils comparent la répartition spatiale de ces maisons avec la répartition qu'elles auraient si elles étaient distribuées aléatoirement parmi l'ensemble de toutes les maisons. Les hypothèses sur la distribution aléatoire dans l'espace permettent d'évaluer la significativité des regroupements de maisons (figure 1.1).

■

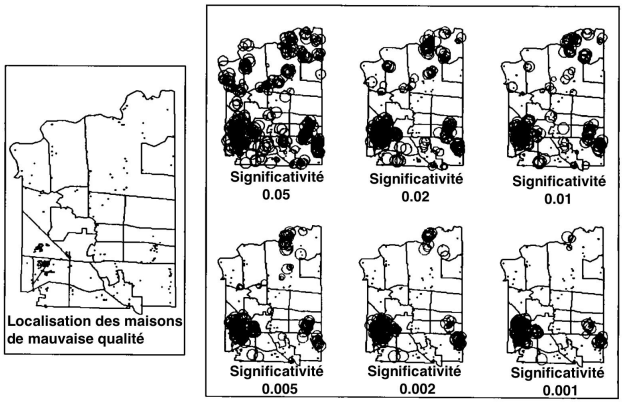


FIGURE 1.1 – Détection de clusters significatifs
Source : FOTHERINGHAM *et al. 1996*

1.1.2 Données continues

En présence de données continues, il existe une valeur pour la variable d'intérêt en tout point du territoire étudié. Les données sont générées de façon continue sur un sous ensemble de $\mathbf{R}^2$. En revanche, ces données sont mesurées uniquement en un nombre discret de points. Il s'agit, par exemple, de la composition chimique du sol (utile à l'industrie minière), de la qualité de l'eau ou de l'air (pour des études sur la pollution), ou encore de diverses variables météorologiques. L'analyse spatiale des données continues, appelée aussi géostatistique, cherche à prédire la valeur d'une variable en un point où elle n'a pas été échantillonnée, ainsi que la fiabilité de cette prédiction. La géostatistique aide également à optimiser le plan d'échantillonnage des données.

R Les principales méthodes permettant d'analyser les données continues sont décrites dans le chapitre 5 : "Géostatistique".

■ **Exemple 1.2 — Prédiction de la pollution.** CHILES *et al. 2005*
Les chercheurs du groupe de travail GeoSiPol (Les pratiques de la géostatistique dans le domaine des sites et sols pollués) prennent en compte la structure de dépendance spatiale entre les données grâce à la technique du *krigeage*. Ils prédisent la quantité de polluant en des lieux où le sol n'a pas été échantillonné et quantifient l'incertitude d'estimation (figure 1.2). ■

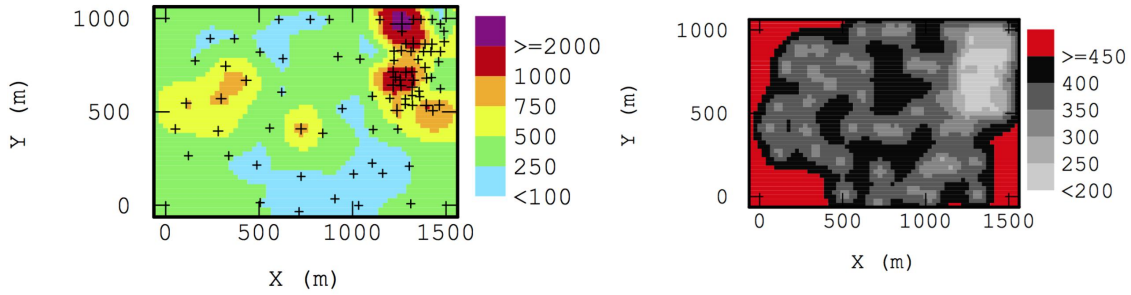


FIGURE 1.2 – Prédiction de la teneur en polluant d'un sol (mg/kg/m2) (à gauche) et écart-type de la prédiction (à droite)
Source : *Manuel GéoSiPol - Mines de Paris :* CHILES *et al. 2005*

1.1.3 Données surfaciques

Pour des données surfaciques, la localisation des observations est considérée comme fixe, mais les valeurs associées sont générées suivant un processus aléatoire. Ces données caractérisent le plus souvent une partition du territoire en zones contiguës, mais elles peuvent également être des points fixes du territoire. Il s'agit, par exemple, du PIB par région, ou du nombre de mariages par mairie. Le terme "surfacique" est donc trompeur, car ces données ne sont pas nécessairement représentées sur une surface. On s'intéresse aux relations entre les **valeurs des observations voisines**. L'analyse spatiale des données surfaciques commence par **définir la structure de voisinage des observations** puis elle **quantifie l'influence qu'exercent les observations sur leurs voisines, et enfin, elle évalue la significativité de cette influence.**

R Les principales techniques d'analyse des données surfaciques sont décrites dans les chapitres 2 : "Codifier la structure de voisinage", 3 : "Indices d'autocorrélation spatiale", ainsi que dans la partie 3.

■ **Exemple 1.3 — Dépendance spatiale locale.** GIVORD *et al. 2016* cherchent à répondre à la question : "Les collèges favorisés sont-ils toujours situés dans un environnement favorisé ?". Les auteurs utilisent pour cela des *Indices locaux d'autocorrélation spatiale* [2]. Ces indices comparent la similarité entre le niveau social d'un collège et celui de son environnement à la similarité qu'ils auraient si les niveaux sociaux des collèges étaient répartis aléatoirement parmi l'ensemble des collèges. Les indices locaux d'autocorrélation spatiale permettent d'identifier les collèges pour lesquels l'influence du milieu social environnant est significative (figure 1.3). ■

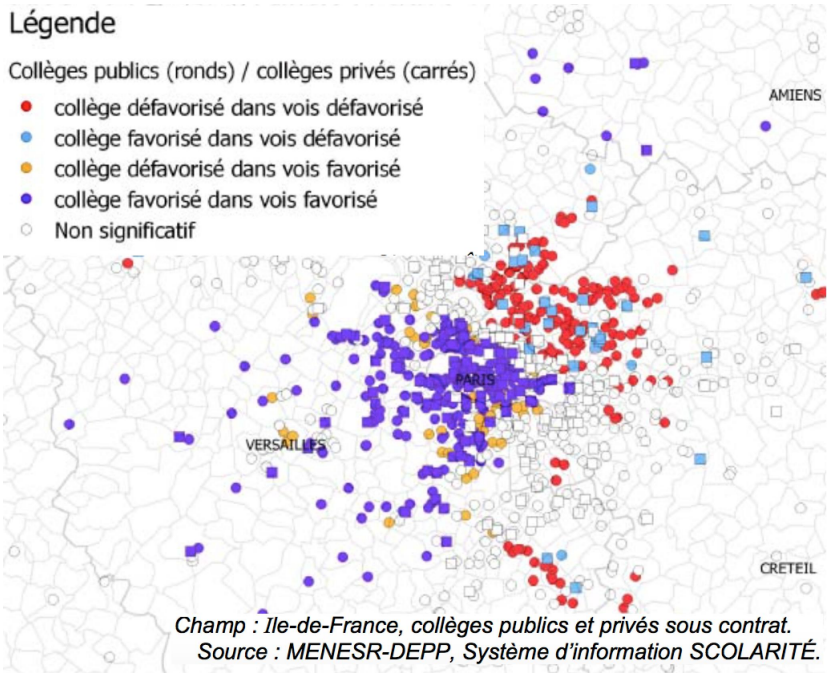


FIGURE 1.3 – Influence du niveau social du voisinage d'un collège sur le niveau social du collège lui-même
Source : GIVORD *et al. 2016*

2. On pourra se reporter au chapitre 3 : "Indices d'autocorrélation spatiale"

Encadré 1.1.1 — Une donnée spatiale peut faire partie de plusieurs catégories. La distinction en trois catégories de données spatiales permet d'orienter l'analyste vers telle ou telle méthode. Il faut néanmoins garder en tête que ces catégories sont perméables et que le choix d'analyser suivant un point de vue un phénomène est lié à l'échelle d'analyse et à l'objectif même de l'étude. Par exemple, une maison est considérée comme un objet ponctuel si on étudie les regroupements dans l'espace significatifs, mais ce peut-être aussi une donnée surfacique si on cherche à connaître la corrélation spatiale entre l'âge des habitants des maisons.

1.2 Notions de sémiologie cartographique

1.2.1 Qu'entend-on par sémiologie cartographique ?

La sémiologie cartographique est l'ensemble des règles qui permettent de transmettre le plus clairement possible une information correcte grâce à une image cartographique. Il est bon d'avoir ces règles en tête avant de passer à la réalisation pratique d'une carte avec le logiciel R. La sémiologie cartographique est un véritable langage destiné à faciliter la communication à l'aide d'outils graphiques appelés variables visuelles. La bonne utilisation de ces variables renforce le message tout en le rendant plus lisible.

Parmi les variables visuelles, on distingue la forme, la texture, la taille de l'objet à représenter, son orientation et sa couleur. Cette dernière peut être associée à des effets de transparence ou afficher un dégradé selon une échelle de valeurs donnée. La dynamique est une variable visuelle plus récente, avec l'apparition notamment des cartes animées.

Les variables visuelles se caractérisent par leur aptitude à mettre en évidence :
- des quantités, souvent représentées par des cercles proportionnels ;
- une hiérarchie, en représentant une série ordonnée de valeurs relatives, par exemple des densités de population ;
- des différences entre entités représentées, par exemple l'industrie et le tourisme ;
- des similitudes, en regroupant en un seul ensemble les différents objets d'un même thème.

Par ailleurs, une combinaison bien appropriée de plusieurs variables visuelles peut renforcer le message.

1.2.2 Objectifs d'une carte

Un graphique permet une appréhension directe et globale d'une information et remplace avantageusement un long tableau. C'est encore plus vrai pour une carte. Son principal intérêt est d'intégrer la dimension spatiale, surtout quand le nombre de territoires est relativement élevé. Ainsi, la carte permet d'un seul coup d'œil de percevoir une information. Grâce à la dimension spatiale, sont prises en compte la situation géographique, la proximité au littoral, à la montagne, aux grandes villes, aux pays voisins, etc. D'où l'importance de rajouter des repères géographiques : régions et pays voisins, noms de ville, fleuves, axes de communication, etc. De plus, la carte est un bon outil de communication. Elle est en effet de compréhension facile : on reconnaît généralement son territoire, et elle est une illustration plaisante. L'évolution technologique des outils de cartographie, gratuits et simples d'accès, permet de réaliser facilement des cartes esthétiques. Cependant, l'esthétisme ne doit pas primer sur la pertinence et encore moins déformer l'information apportée par la carte.

1.2.3 À chaque type de données, sa variable visuelle

La première question à se poser est de savoir ce que l'on veut représenter. En effet, pour représenter une variable en volume ou un chiffre absolu, on utilise des ronds proportionnels ; pour

des ratios, densités, évolutions, parts et typologie, on utilise une carte en aplats de couleurs ; les données bilocalisées ou les flux sont illustrés par des oursins, flèches proportionnelles ou résultantes vectorielles. Enfin, la localisation, par exemple d'équipements, se fait par des cartes à symboles.

Dans le cas d'une carte en aplats de couleurs (appelée aussi analyse en classes ou carte choroplète), les valeurs positives sont dans des teintes chaudes (rouge, orange) alors que les valeurs négatives sont généralement dans des teintes froides (bleu, vert). Par ailleurs, à une hiérarchie de valeurs correspond un dégradé de couleurs dont les couleurs les plus foncées (ou les plus claires) correspondent aux valeurs extrêmes.

Des règles existent également pour la discrétisation des données, c'est-à-dire la manière dont les observations sont regroupées en classes. Le nombre de classes se calcule en fonction du nombre d'observations. Différentes théories existent pour déterminer le nombre optimal. Selon la règle de Sturges par exemple, il est égal à $1+3,3*log_{10}(N)$, où N est le nombre d'observations.

En pratique :
— pour moins de 50 observations : 3 classes ;
— pour 50 à 150 observations : 4 classes ;
— pour plus de 150 observations : 5 classes.

La forme de la distribution des données nous aide aussi dans ce choix. Ainsi, on rajoute une classe en cas de présence de valeurs négatives et positives. Une fois le nombre de classes déterminé, une méthode de regroupement doit être choisie. Plusieurs méthodes existent, chacune d'elles présentant des avantages et des inconvénients.

— **La méthode des quantiles** : elle consiste à utiliser le même nombre de valeurs par classe. Elle produit une carte harmonieuse et facile à lire, les couleurs de la légende se répartissant à parts égales. Cependant, elle ne s'adapte pas toujours à la distribution des données.
— **La méthode des classes de même amplitude** : elle consiste à découper l'intervalle de valeurs en plages de même longueur. Cette méthode est simple à comprendre mais s'adapte très rarement à la distribution ; certaines classes peuvent ne contenir aucune valeur.
— **Les méthodes de Jenks et k-means** : elles visent à créer des classes homogènes en maximisant la variance entre les classes et en minimisant la variance au sein de chacune d'entre elles. Ces méthodes, contrairement aux deux précédentes, s'adaptent parfaitement aux données en éliminant les effets de seuil. Cependant, le temps de calcul de la méthode de Jenks peut être très long si les observations sont nombreuses. Pour cette raison, on peut utiliser la méthode k-means dont le calcul est plus rapide même avec un nombre élevé d'observations. Celle-ci peut cependant être instable, en donnant des classes différentes pour un même jeu de données. On gère ce problème en répétant la k-means plusieurs fois pour garder la meilleure répartition.
— **La méthode de l'arrangement manuel** : elle consiste à fixer soi-même les bornes des classes. Elle est utile pour faire apparaître des valeurs significatives (borne à zéro ou autour de zéro, moyenne…) ou pour améliorer à la marge le positionnement de certains seuils en fonction de la distribution locale. Elle permet également de rendre des cartes comparables entre elles en fixant des bornes identiques de classes. Cette méthode nécessite d'analyser au préalable la distribution des données, en utilisant dans un premier temps la méthode de Jenks ou de k-means pour avoir des classes homogènes puis en ajustant les bornes des classes manuellement pour éviter les effets de seuil.

1.2.4 Quelques conseils

— **Un message simple par carte**. Une carte est souvent difficilement compréhensible quand elle comporte trop d'informations. Par exemple, aucun message ne se dégage de la carte présentée en figure 1.4 car elle est trop compliquée. D'où la règle élémentaire de faire simple

pour être efficace. Pour ce faire, le nombre de variables à représenter sur une même carte doit être limité.

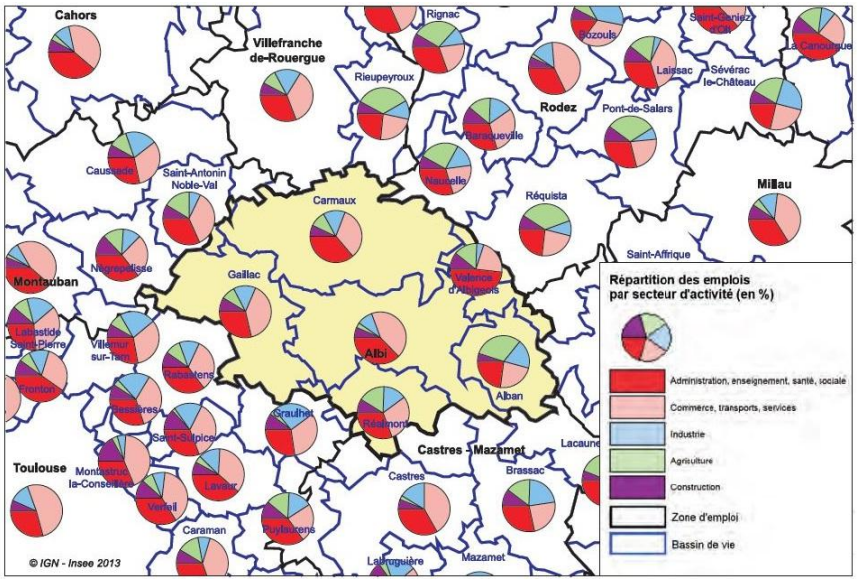


FIGURE 1.4 – Répartition des emplois par secteur d'activité dans les bassins de vie
Source : *Insee, recensement de la population 2010*

— **Faire figurer les informations de base**. Une carte doit impérativement comporter un titre informatif (le plus souvent associé à un sous-titre descriptif), une mention du zonage représenté, une légende, une source et un copyright. L'échelle, un logo ou la flèche Nord peuvent éventuellement y figurer.
— **Ne pas présenter le territoire comme une île**. Il est souhaitable de donner des éléments d'environnement au lecteur pour qu'il situe le territoire représenté ; par exemple, les départements ou régions limitrophes, des éléments de topographie comme la mer ou le réseau routier. Dans la figure 1.5, il aurait été judicieux de représenter les communes des départements environnants, notamment Dijon au nord ou Lyon au sud pour pouvoir illustrer le titre qui n'est pas très explicite.

  Par ailleurs, il peut être intéressant d'élargir les analyses effectuées sur le territoire à l'environnement alentour, à la condition que le territoire d'intérêt ressorte bien comme dans la figure1.6 (contour vert foncé et trame vert clair). L'analyse élargie permet ici de situer le dynamisme démographique de Toulouse par rapport à celui de Bordeaux et de mieux comprendre l'importance du système urbain languedocien, dans la continuité de celui du couloir rhodanien.
— **Des cartes comparables**. Lorsque deux cartes illustrant le même territoire avec les mêmes variables visuelles sont disposées côte à côte ou l'une en dessous de l'autre, le lecteur est incité à faire des comparaisons. Pour faciliter cette opération, les deux cartes doivent avoir une légende harmonisée (mêmes classes, cercles ou flèches) et une même échelle avec un zoom identique. Dans les cartes de la figure 1.7, les légendes harmonisées permettent de comparer l'évolution annuelle de la population sur les deux périodes 1982-2011 et 2006-2011.
— **Choisir son indicateur : parts ou effectifs ?** L'analyse en classes est utilisée pour représenter une sous-population en valeur relative (ou part) ou une évolution. Elle est prohibée pour la représentation d'effectifs ou de volumes car elle pourrait induire le lecteur à interpréter la carte de manière erronée. L'œil établirait en effet une correspondance entre le volume repré-

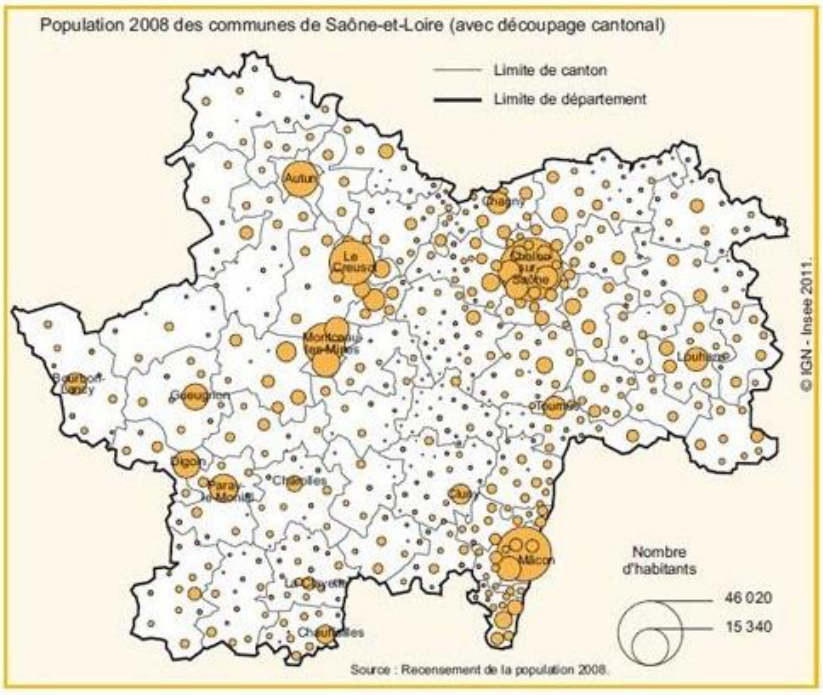


FIGURE 1.5 – Plusieurs villes moyennes
Source : *Insee, Recensement de la Population 2008*

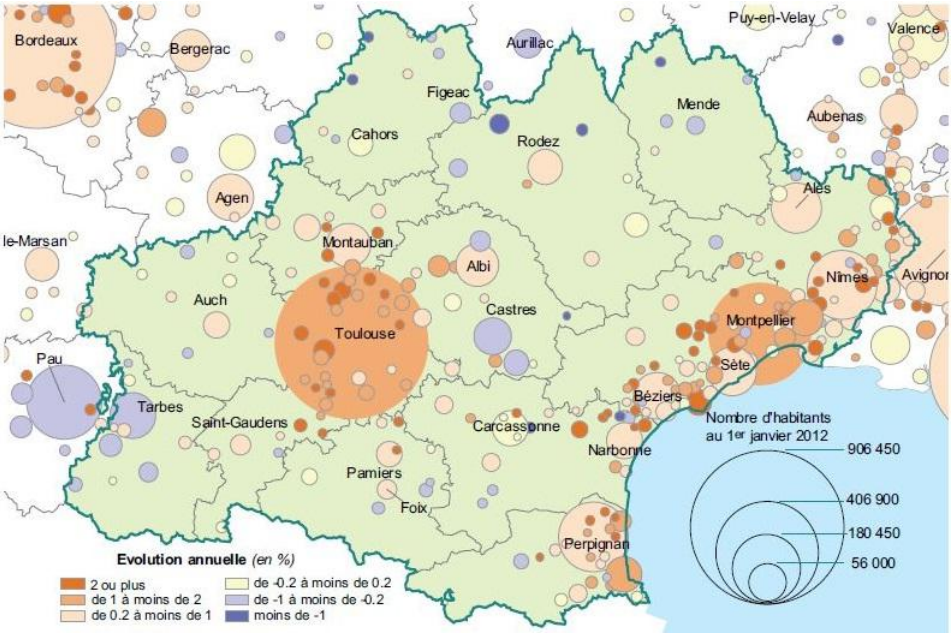
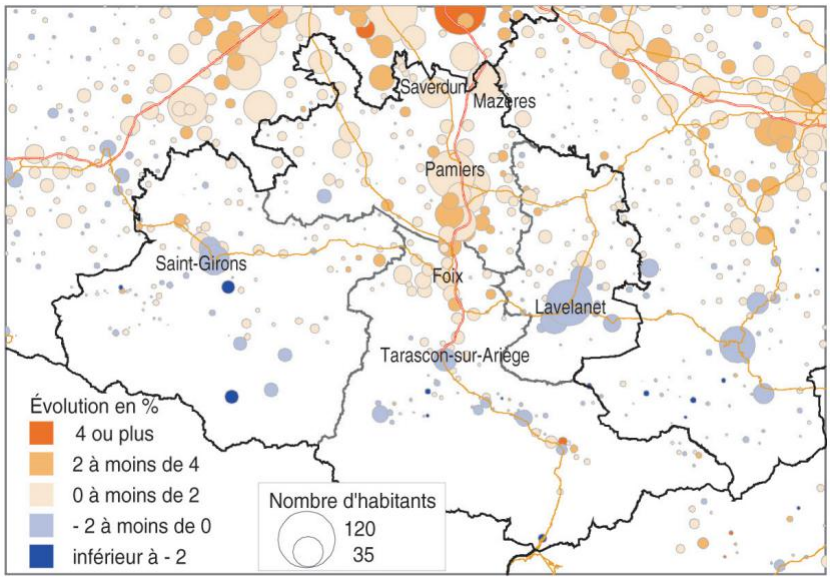
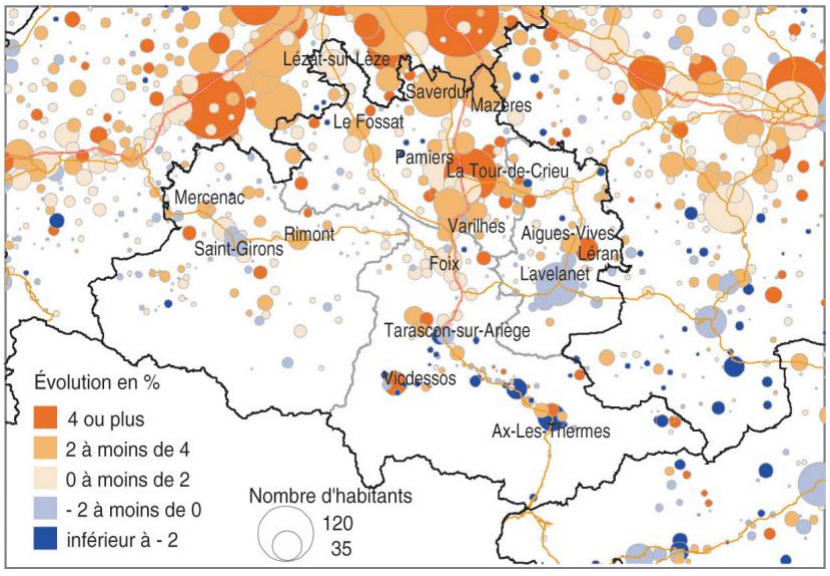


FIGURE 1.6 – Un système urbain monocentré autour de Toulouse et polycentré sur le littoral
Source : *Insee, recensements de la population 2007 et 2012*



(a) Évolution annuelle de la population des communes entre 2006 et 2011



(b) Évolution annuelle de la population des communes entre 1982 et 2011

FIGURE 1.7 – Évolution annuelle moyenne de la population des communes de Basse-Ariège
Source : *Insee, Recensements de la Population 1982, 2006 et 2011*

senté et la surface du territoire colorié. Ainsi, une analyse en classes sur le nombre d'habitants par commune induirait une surestimation visuelle de la population d'Arles, commune la plus étendue de France. Par ailleurs, une analyse en classes seule peut parfois être trompeuse car des pourcentages élevés peuvent concerner des petits effectifs. C'est pourquoi, il est parfois nécessaire de combiner ce type d'analyse avec une analyse en ronds proportionnels portant sur les effectifs. Selon le message que l'on veut faire passer, on choisira de colorier des ronds avec une analyse en classes (figure 1.7) ou de plaquer des ronds sur une analyse en classes (figure 1.8). Dans le cas de ronds coloriés, l'œil est davantage attiré par la taille des ronds et dans l'autre cas, l'œil sera d'abord attiré par les couleurs les plus foncées de l'analyse en classes.

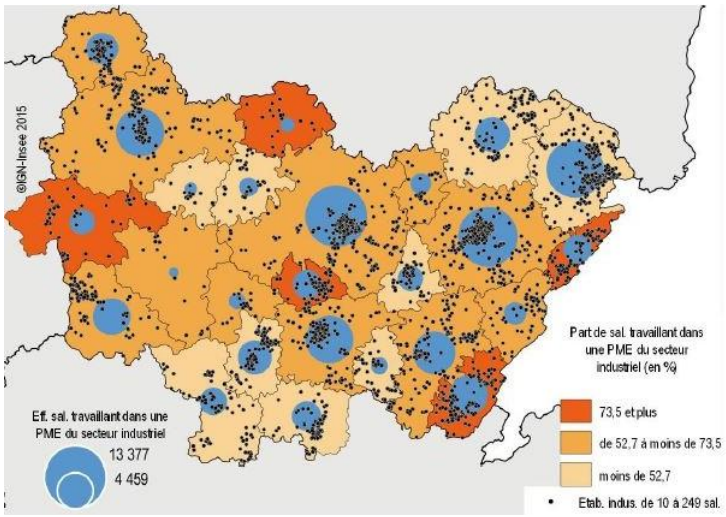


FIGURE 1.8 – Répartition des salariés travaillant dans une PME du secteur industriel
Source : *Insee, Connaissance Locale de l'Appareil Productif 2012*

1.3 Éléments de cartographie avec R

Les données géolocalisées peuvent être agrégées à une échelle géographique plus ou moins grande. On peut ensuite les cartographier de différentes façons. Nous décrirons dans cette section comment appréhender simplement la cartographie avec R, et quelques packages appropriés. De nombreux packages permettent de représenter des données spatiales ; ceux que nous mettrons en œuvre dans ce manuel sont :

— *sp* : package de base définissant les objets spatiaux ;
— *rgdal* : import/export d'objets spatiaux ;
— *rgeos* : manipulation de la géométrie ;
— *cartography* : réalisation de cartes d'analyse.

Nous présenterons également le package *sf* qui regroupe l'ensemble des fonctions des packages *sp*, *rgdal* et *rgeos*.

1.3.1 Manipulation d'objets spatiaux

Points, Polygones, Lignes

Le package *sp* permet de créer ou de convertir en `objet` *sp* différentes géométries : des points, des lignes, des polygones ou encore des grilles. En général, chaque `objet` *sp* est composé de différentes parties : les slots. Chaque slot contient une information particulière (coordonnées géographiques, table d'attributs, système de coordonnées, étendue spatiale, etc.)

L'accès à un slot d'un `objet` *sp* se fera à l'aide de l'opérateur `@` (`objet@slot`).

Les objets spatiaux peuvent être abordés sous différentes formes. La première correspond à des **points**, c'est-à-dire un ensemble de points géoréférencés.

```
library(sp)

# contenu d'une table communale contenant les coordonnées des mairies
# en WGS84 (latitude/longitude)

head(infoCom)
##         nom_commune latitude longitude préfecture
##               <chr>    <dbl>     <dbl>      <chr>
## 1  Faches-Thumesnil 50.58333  3.066667      Lille
## 2             Lille 50.63333  3.066667      Lille
## 3          Lezennes 50.61667  3.116667      Lille
## 4             Lille 50.63333  3.066667      Lille
## 5           Ronchin 50.60000  3.100000      Lille
## 6 Villeneuve-d'Ascq 50.68333  3.141667      Lille

# Transformation en objet spatial

communes <- SpatialPoints(coords=infoCom[,c(2,3)])

#Visualisation des slots disponibles

slotNames(communes)
##[1] "coords"      "bbox"        "proj4string"

# Connaitre l'étendue spatiale

communes@bbox  # ou bbox(communes)
##               min       max
##latitude  50.000000 51.083333
##longitude  2.108333  4.183333
```

On peut aussi représenter graphiquement cet objet *via* l'instruction graphique classique `plot` (illustration en figure 1.9).

```
plot(communes)
```

Notre objet spatial peut également posséder une table d'attributs décrivant les objets géographiques qu'il contient. L'objet appartient alors à la classe des `SpatialPointsDataFrame` :

```
#Ajout de la table d'attributs
```

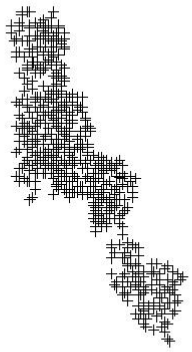


FIGURE 1.9 – Communes du Nord de la France
Source : *Insee*

```
nord<-  SpatialPointsDataFrame(coords=infoCom[,c(2,3)],data =infoCom[,c
    (1,4)])
```

On accède à cette table d'attributs *via* le nouveau slot créé `@data` :

```
nord@data
##          nom_commune préfecture
##                <chr>      <chr>
## 1   Faches-Thumesnil      Lille
## 2              Lille      Lille
## 3           Lezennes      Lille
## 4              Lille      Lille
## 5            Ronchin      Lille
## 6  Villeneuve-d'Ascq      Lille
## 7       La Madeleine      Lille
## 8              Lille      Lille
## 9            Comines      Lille
## 10         Deulemont      Lille
## # ... with 611 more rows
```

La création de **polygones géoréférencés**, bien qu'un peu plus complexe, suit la même logique.

En premier lieu nous allons créer des polygones simples à l'aide des coordonnées des sommets :

```
# Création des séries de coordonnées
x1 <- c(439518.5, 433091.8, 455774.1, 476566.1, 476944.2, 459554.4,
    439518.5)
y1 <- c(8045280, 8031293, 8018439, 8026756, 8044902, 8054731, 8045280)
        c1 <- data.frame(x1, y1)
x2 <- c(444929.2, 417667.9, 501837.1, 499792.5, 444929.2)
y2 <- c(8121306, 8078029, 8067465, 8109039, 8121306)
```

```
        c2 <- data.frame(x2, y2)
x3 <- c(456530.1, 450481.5, 472785.8, 476566.1, 456530.1)
y3 <- c(8101608, 8089510, 8087620, 8099717, 8101608)
        c3 <- data.frame(x3, y3)

# création des polygones
p1 <- Polygon(coords = c1, hole = F)
p2 <- Polygon(coords = c2, hole = F)
p3 <- Polygon(coords = c3, hole = T)
```

Le paramètre `hole` sert à identifier les polygones représentant des trous à l'intérieur d'autres polygones.

Ces objets possèdent 5 slots, dont :
- `@labpt` qui donne les coordonnées du centre ;
- `@hole` qui dit s'il s'agit d'un trou ;
- `@coords` qui permet de récupérer les coordonnées des sommets.

Ils peuvent ensuite être assemblés en polygones multiples :

```
P1 <- Polygons(srl = list(p1), ID = "PolygA")
P2 <- Polygons(srl = list(p2, p3), ID = "PolygB")
```

Ainsi le polygone *P1* sera composé de *p1* et *P2* sera *p2* avec un trou au centre défini par *p3*

Ils possèdent encore 5 slots différents, dont :
- `@Polygons` qui donne la liste des polygones ayant servi à sa création ;
- `@ID` qui donne les identifiants donné au polygone.

On spatialise ensuite cet ensemble de polygones pour en faire un unique objet spatial :

```
SP <- SpatialPolygons(Srl = list(P1, P2))
```

Notre objet spatial se structure donc de la manière suivante : le `SpatialPolygons` contient une liste de deux `polygons` (polygones multiples) contenant chacun une liste de `Polygons` (polygones simples), lesquels contiennent les coordonnées qui les délimitent. Ainsi, pour accéder aux coordonnées du premier polygone simple contenu dans le second polygone multiple, nous devons écrire :

```
SP@polygons[[2]]@Polygons[[1]]@coords
```

```
##          x2      y2
## [1,] 444929 8121306
## [2,] 499793 8109039
## [3,] 501837 8067465
## [4,] 417668 8078029
## [5,] 444929 8121306
```

Pour ajouter une table d'attributs à notre objet géographique, il suffit de créer un `dataframe` contenant autant de lignes que de polygones multiples dans notre objet. Les lignes doivent être triées dans le même ordre que les polygones et chaque ligne identifiée par le même identifiant.

```
Info <- c("Simple", "Hole")
Value <- c(342, 123)
mat <- data.frame(Info, Value)
rownames(mat) <- c("PolygA", "PolygB")
```

```
SPDF<-  SpatialPolygonsDataFrame(Sr = SP, data= mat )
```

Un nouveau slot est rajouté, `@data`, pour récupérer la table des attributs. On peut représenter graphiquement cet objet, ce qui donne la figure 1.10.

```
plot(SPDF,col =c("lightgrey","black"))
```

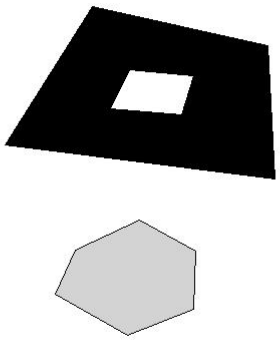


FIGURE 1.10 – Polygones créés

On peut construire des objets de type **lignes géoréférencées** de la même manière que celle présentée précédemment pour les polygones. Cette fois-ci, ce seront les fonctions `SpatialLines` et `SpatialLinesDataFrame` qui serviront.

Ainsi introduits, nos fonds de cartes communaux, départementaux, etc, seront des objets de type `SpatialPolygons(DataFrame)`, nos fonds routiers ou cours d'eau de type `SpatialLines(DataFrame)` et nos fonds d'aéroports ou des mairies de type `SpatialPoints(DataFrame)`.

Travail sur une couche vectorielle

La grande majorité du temps, nous ne créons pas d'objets géographiques de toute pièce mais manipulons des objets déjà existants. Plusieurs packages permettent d'importer ou exporter des objets géographiques. Le plus simple et complet demeure *rgdal* qui permet de lire et manipuler un très grand nombre de formats. Le format vectoriel le plus répandu est l'"ESRI ShapeFile", qui fournit un fond de carte à travers 5 fichiers devant être présents côte à côte dans un même dossier (.shp, .shx, .dbf, .prj, .cpg). Tous ces fichiers portent le même nom, seule l'extension diffère.

Pour importer le fond de carte, on utilise la fonction `readOGR` :

```
library(rgdal)
comr59<-  readOGR(dsn = "Mes Documents\\fonds", layer = "comr59", verbose =
    FALSE)
```

Les paramètres de la fonction `readOGR` sont :

- `dsn` : chemin du dossier où sont stockés les fichiers ;
- `layer` : nom du fichier (sans extension).

On obtient alors un objet R de type `SpatialPolygonsDataFrame` (exemple en figure 1.11).

```
class(comr59)
## [1] "SpatialPolygonsDataFrame"
## attr(,"package")
## [1] "sp"
plot(comr59)
```

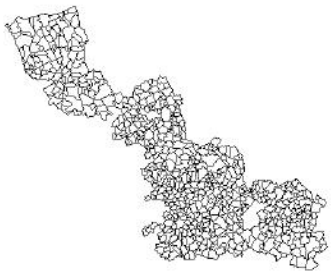


FIGURE 1.11 – Fond communal Nord
Source : *Insee*

`readOGR` permet d'importer une large palette de formats cartographiques. Pour un fond issu de MapInfo, la syntaxe change peu :

```
comr59MI<- readOGR(dsn= "Mes documents\\fonds\\comr59.tab", layer="comr59",
    verbose=FALSE)
```

Comme pour les ShapeFiles, le format MapInfo se compose d'un certain nombre de fichiers qui doivent être tous présents dans le même dossier, avec un nom identique mais des extensions différentes. Dans ce cas, le `dsn` pointe jusqu'au fichier .tab, et le `layer` prend le nom des fichiers du fond.

Pour sélectionner un sous-ensemble de notre carte, on se réfère au dataframe associée *via* le slot `@data`. Ainsi, pour sélectionner les communes de superficie supérieure à 200 km^2 :

```
comr59_etendue<-  comr59[comr59@data$surf_m2>20000000, ]
```

Pour visualiser cette sélection, on superpose les 2 objets en colorant la sélection en gris (résultat visible en figure 1.12) :

```
plot(comr59)
plot(comr59_etendue,col ="darkgrey",add =TRUE)
```

Le paramètre `add=TRUE` permet de superposer les 2 fonds.

Pour sauvegarder notre nouveau fond cartographique, on utilise la fonction `writeOGR` qui prend comme paramètres :

- `obj` : objet R à exporter ;
- `dsn` : chemin du dossier de sauvegarde ;
- `layer` : nom commun des fichiers (sans extension) ;
- `driver` : format d'exportation de l'objet.

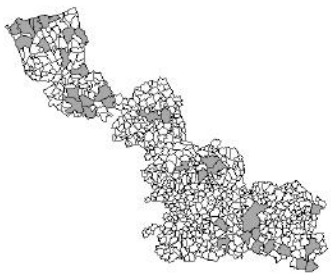


FIGURE 1.12 – Fond communal Nord étendu
Source : *Insee*

L'ensemble des formats possibles est fourni par la fonction `ogrDrivers()`.
Ainsi pour exporter notre sélection au format ShapeFile :

```
writeOGR(comr59_etendue,dsn="Mes documents\\fonds",layer="comr59_etendue",
driver="ESRI Shapefile")
```

Au format MapInfo :

```
writeOGR(comr59_etendue, dsn="Mes documents\\fonds\\comr59_etendue_MI.tab",
layer="comr59_etendue_MI", driver="MapInfo File")
```

1.3.2 Réalisation de cartes statistiques

Le système de projection

Les données spatiales sont toujours associées à un système de projection. Celui-ci est identifiable par le slot `@proj4string`.

```
comr59@proj4string
## CRS arguments:
##  +proj=longlat +datum=WGS84 +no_defs +ellps=WGS84 +towgs84=0,0,0
```

On peut affecter un système de projection à un objet créé. Si on reprend nos polygones précédents, ils n'ont par défaut aucun système de projection associé. Pour signifier qu'ils sont en WGS84 :

```
SPDF@proj4string<- CRS( "+proj=longlat +datum=WGS84 +ellps=WGS84")
```

Le standard EPSG permet également d'identifier les système de projection par un code unique. Il est de 4326 pour le WGS84. Dans ce cadre, l'affectation précédente pourrait se coder :

```
SPDF@proj4string<- CRS( "+init=epsg:4326")
```

L'ensemble des correspondances des codes EPSG peut être obtenu en exécutant `make_EPSG()`.
Ainsi pour le Lambert 93 (utilisé entre autres par l'IGN) le code EPSG est le 2154.

Si maintenant on souhaite reprojeter un objet géographique dans un nouveau système de coordonnées, on utilise la fonction `spTranform()`.

```
comr59_l93<-  spTransform(comr59, CRSobj=CRS("+init=epsg:2154"))
```

Cette reprojection est nécessaire notamment pour pouvoir superposer deux fonds qui ne possèdent pas le même système de coordonnées.

Pour réaliser très simplement des cartes, nous présentons le package *cartography* qui en plus de sa facilité de prise en main est relativement complet dans ses possibilités.

Cartes en symboles proportionnels :
La cartographie de données de stocks (comme une population, un nombre d'équipements...) se fait à l'aide de symboles proportionnels à la grandeur représentée. Le plus commun est le rond, mais on peut imaginer tout autre symbole. Le code ci-dessous permet d'obtenir la figure 1.13.

```
library(rgdal)
library(cartography)
metr_nice <- readOGR(dsn="~\\fonds",layer="metr_nice",verbose=F)

# Table des données de population
head(donnees_communales)
##   CODGEO                    LIBGEO REG DEP P13_POP
## 1  01001  L'Abergement-Clémenciat   84  01     767
## 2  01002    L'Abergement-de-Varey   84  01     236
## 3  01004        Ambérieu-en-Bugey   84  01   14359
## 4  01005      Ambérieux-en-Dombes   84  01    1635
## 5  01006                  Ambléon   84  01     108
## 6  01007                 Ambronay   84  01    2503

#tracé du fond de carte
plot(metr_nice)

#ajout de l'analyse
propSymbolsLayer(spdf=metr_nice, df=donnees_communales, spdfid = "Codgeo",
                         dfid = "CODGEO", var= "P13_POP", col="salmon",
            symbols="circle", legend.pos="right")

#habillage de la carte
layoutLayer(title = "Population de Nice métropole",
            author = "INSEE", sources = "Recensement 2013",
            scale = NULL, north = TRUE)
```

Les différents paramètres de la fonction sont :
- `spdf` : le SpatialPolygonsDataFrame ;
- `df` : le dataframe contenant les données à analyser ;
- `spdfid` : identifiant de la maille cartographique (dans le slot `@data`) ;
- `dfid` : identifiant de ligne dans le dataframe. Doit correspondre avec le précédent ;
- `var` : variable du dataframe à analyser.

D'autres paramètres existent et peuvent être listés dans l'aide de la fonction.

Cartes choroplèthes :
Pour la représentation des taux, on utilise des cartes en aplats de couleur ou choroplèthes. La variable est répartie dans des classes et un dégradé de couleurs schématise la croissance des valeurs (voir figure 1.14).

```
plot(metr_nice)
```

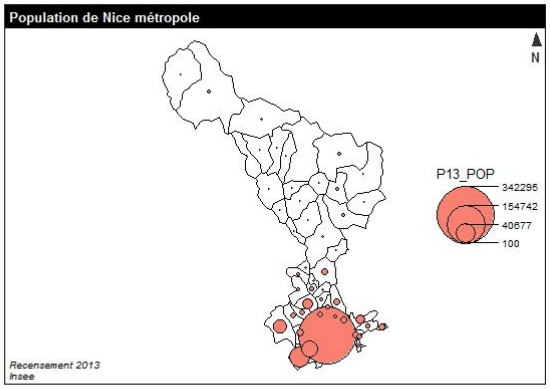


FIGURE 1.13 – Symboles proportionnels
Source : *Insee, Recensement de la population 2013*

```
choroLayer(spdf=metr_nice,df =donnes_communales4, spdfid = "Codgeo",
           dfid = "CODGEO",var = "TCHOM", nclass=4, method="fisher-jenks",
           legend.pos="right")

layoutLayer(title=  "Taux de chomage des communes de Nice metropole",
            author = "Insee", sources = "Recensement 2013",
            scale = NULL, north = TRUE)
```

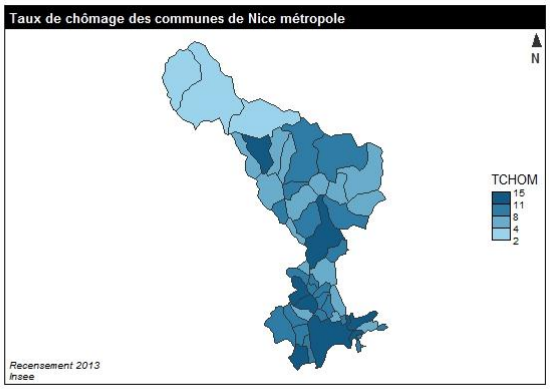


FIGURE 1.14 – Carte choroplèthe
Source : *Insee, Recensement de la population 2013*

La classification se réalise soit en spécifiant le nombre de classes (`nclass`) et la méthode de découpage (`method` qui permet de choisir parmi les méthodes présentées en section 1.2) ou en donnant un vecteur de bornes (`breaks`).

Autres fonctions cartographiques :

— *propSymbolsChoroLayer* : il s'agit d'un mélange entre les symboles proportionnels et les cartes choroplèthes (pour représenter simultanément un nombre de chômeurs et un taux de chômage par exemple) ;
— *typoLayer* : pour représenter une typologie en spécifiant une variable qualitative et un vecteur de couleur de même longueur que le nombre de modalités ;
— *gradLinkLayer* : pour représenter des flux ou des liens.

D'autres packages permettent de réaliser des cartes statistiques sous R. On peut citer entre autres :

— *RgoogleMaps* : réaliser des cartes en utilisant des rasters routiers ou satellites GoogleMaps ;
— *leaflet* : réaliser des cartes interactives avec raster OpenStreetMap pouvant être insérées dans des pages Web voire du RShiny.

1.3.3 *sf* : l'avenir du traitement des données spatiales sous R

Comme nous l'avons vu précédemment, jusqu'à présent le traitement des données cartographiques se faisait à travers trois packages principaux sous R :

— *sp* pour l'implémentation des classes de type spatiales ;
— *rgdal* pour les bibliothèques d'entrée/sortie ;
— *rgeos* pour les opérations sur les objets géométriques.

Depuis peu, il existe un package unique, nommé *sf*, qui regroupe l'ensemble des fonctionnalités de ces 3 packages réunis. Il fournit aux utilisateurs une classe unique pour la manipulation de l'ensemble des objets spatiaux. Dans ce chapitre, nous présentons rapidement les principales fonctionnalités du package *sf*. Pour aller plus loin dans la compréhension de ce package, la manipulation des géométries riches ou encore la gestion des entrées/sorties, nous invitons le lecteur à consulter les différentes vignettes mises à disposition avec le package sur le site du CRAN. En effet, ce package n'est pas encore compatible avec l'ensemble des packages d'analyse spatiale présentés dans ce manuel qui sont le plus souvent construits à partir de *sp*, *rgdal* et *rgeos*.

Il est notable que les entrées/sorties sont bien plus rapides avec *sf* qu'avec *rgdal*.

Les objets de classe *sf* se définissant comme des `data.frame` agrémentés d'attributs géométriques, la manipulation des objets géographiques en est simplifiée et se fait nativement comme une table quelconque sous R. Concrètement le package définit trois classes d'objets différents :

— `sf` : un `data.frame` avec des attributs spatiaux ;
— `sfc` : la colonne du `data.frame` stockant les données géométriques ;
— `sfg` : la géométrie de chaque enregistrement.

Ainsi un objet spatial sera représenté comme en figure 1.15.

L'importation de fond cartographique existant est simplifiée sous *sf*. Elle se fait ainsi :

```
library(sf)
depf <- st_read("J:/CARTES/METRO/An15/Shape/Depf_region.shp")
```

Il est à noter qu'il n'est pas nécessaire de spécifier le driver d'importation : `st_read` s'adapte automatiquement au format du fichier en entrée. La fonction est compatible avec la grande majorité des formats cartographiques courants (ESRI-Shapefile, MapInfo, PostGIS, etc.). On peut cartographier facilement les données spatiales avec la fonction `plot` (figures 1.16 et 1.17).

L'exportation de fond est tout aussi simple :

```
st_write(depf, "U:/fond_dep.shp")
```

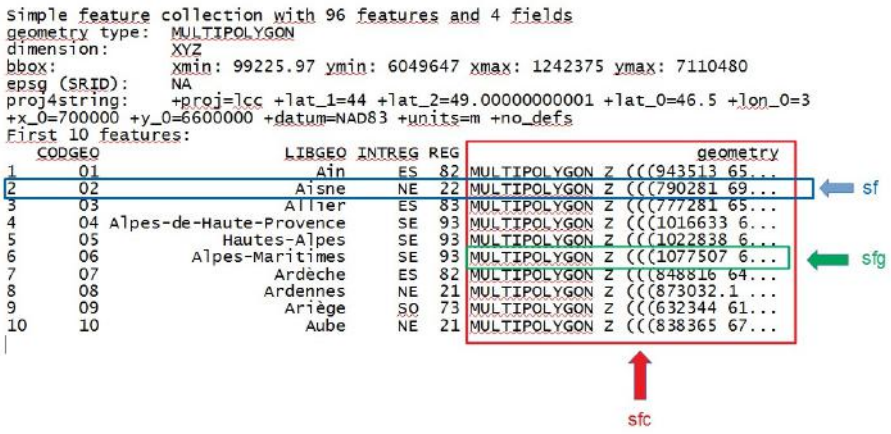


FIGURE 1.15 – Représentation d'un objet spatial avec le package *sf*

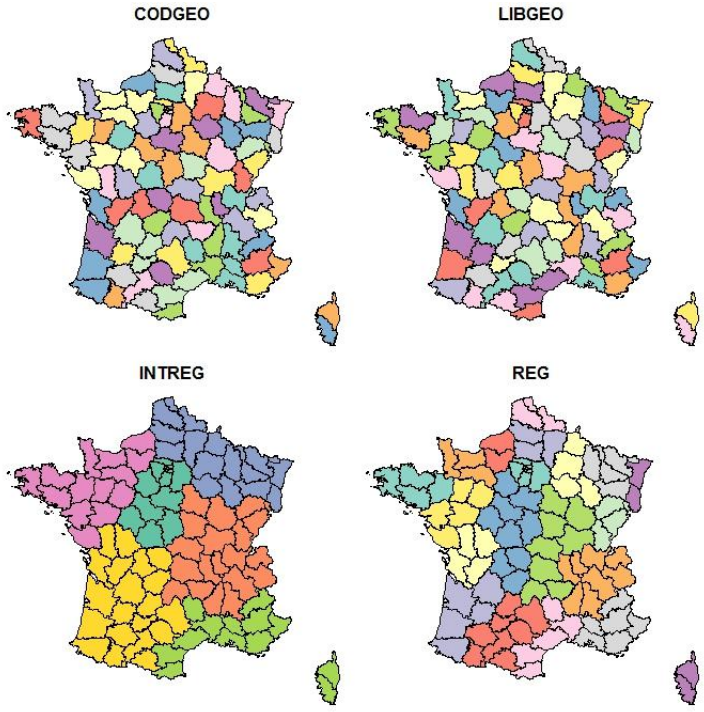


FIGURE 1.16 – Carte obtenue avec le code : `plot(depf)`

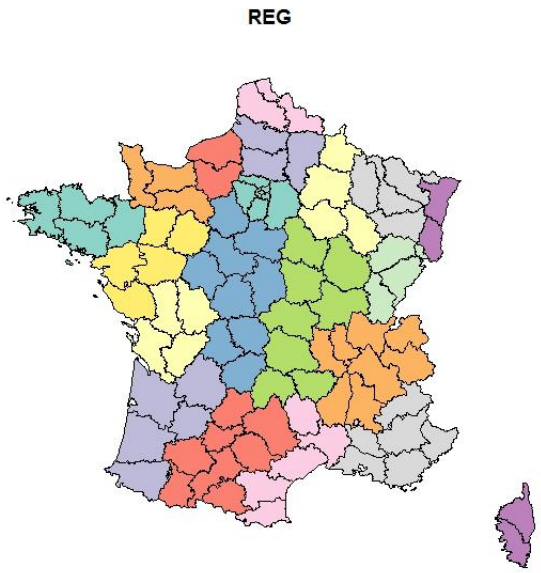


FIGURE 1.17 – Carte obtenue avec le code : `plot(depf["REG"])`

Plus généralement, le package *sf* propose un ensemble d'opérateurs de données spatiales tous préfixés par `st_` et présentés en figure 1.18 :

```
##  [1] "st_agr<-.sf"          "st_agr.sf"            "st_as_sf.sf"
##  [4] "st_bbox.sf"           "st_boundary.sf"       "st_buffer.sf"
##  [7] "st_cast.sf"           "st_centroid.sf"       "st_convex_hull.sf"
## [10] "st_coordinates.sf"    "st_crs<-.sf"          "st_crs.sf"
## [13] "st_difference.sf"     "st_geometry<-.sf"     "st_geometry.sf"
## [16] "st_intersection.sf"   "st_is.sf"             "st_line_merge.sf"
## [19] "st_make_valid.sf"     "st_polygonize.sf"     "st_precision.sf"
## [22] "st_segmentize.sf"     "st_set_precision.sf"  "st_simplify.sf"
## [25] "st_split.sf"          "st_sym_difference.sf" "st_transform.sf"
## [28] "st_triangulate.sf"    "st_union.sf"          "st_voronoi.sf"
## [31] "st_zm.sf"
```

FIGURE 1.18 – Ensemble des opérateurs présents dans le package *sf*

Le package *sf* est également totalement intégré dans l'environnement `tidyverse` et ainsi se conjugue parfaitement avec les fonctionnalité du package *dplyr* :

```
library(dplyr)
depf<- left _join(depf,pop_dep,by="CODGEO")
```

ou encore le code ci-dessous, illustré dans la figure 1.19.

```
library(dplyr)
depf %>%
  mutate(
```

```
  area = st_area(.), # on crée la nouvelle variable sur la superficie
) %>%
group_by(REG) %>%
summarise(mean_area = mean(area)) %>%
plot
```

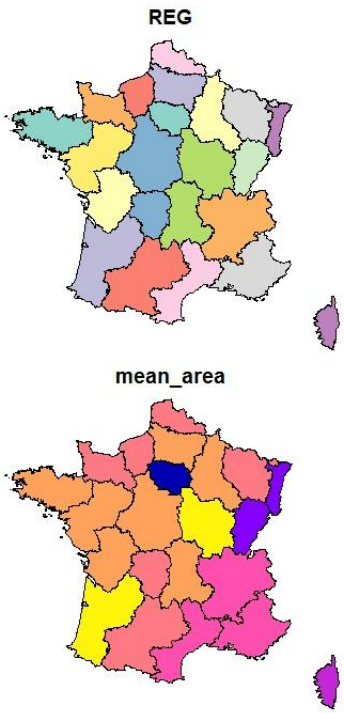


FIGURE 1.19 – Carte réalisée grâce aux packages *sf* et *dplyr*

La plupart des packages liés au traitement de données géographiques se sont adaptés à cette nouvelle classe d'objets. Certains, comme *spdep*, sont en phase de test de leur adaptation et nécessitent ainsi encore le recours au package *sp*.

Concernant *cartography*, l'adaptation est effective depuis sa version 2.0 et la syntaxe a évolué :

```
choroLayer(x, spdf, spdfid,df , dfid,var , ...)
```

où `x` est un objet de type `sf`. S'il est renseigné, les objets `spdf`, `spdfid`, `df` et `dfid` sont ignorés car l'ensemble des ces renseignements sont inclus dans l'objet `x`.

1.3.4 De la surface au point, et réciproquement

Une particularité des données surfaciques est qu'elles peuvent consister en une partition de l'ensemble du territoire ou en un ensemble de points de référence aux coordonnées géographiques distinctes. On peut cependant passer facilement d'une représentation à l'autre :

— les polygones de Voronoï permettent de créer une partition du territoire à partir de points de référence ;
— utiliser le centroïde d'une aire permet de passer d'une partition du territoire à un ensemble de points.

Définition 1.3.1 — Polygone de Voronoï associé au point x_i**.** Il s'agit de la région de l'espace qui est plus proche de x_i que de tout autre point de l'ensemble d'étude $\mathbf{x}$:

$$C(x_i|\mathbf{x}) = \left\{ u \in \mathbb{R}^2 : ||u - x_i|| = \min_j ||u - x_j|| \right\} \tag{1.1}$$

Encadré 1.3.1 — Des polygones très utilisés. Les polygones de Voronoï font partie des structures géométriques les plus utilisées par la communauté scientifique. D'après AURENHAMMER 1991, trois grandes raisons expliquent cet intérêt. La première est que les polygones de Voronoï sont directement observables dans la nature (dans les arrangements cristallins, par exemple). Deuxièmement, ils sont l'une des constructions les plus fondamentales définies par un ensemble discret de points : ils présentent de très nombreuses propriétés mathématiques, et sont reliés à plusieurs autres structures géométriques fondamentales. Enfin, les polygones de Voronoï permettent de simplifier un grand nombre de problèmes algorithmiques. Le polygone de Voronoï associé à un point est souvent considéré comme sa "zone d'influence".

Historiquement, Gauss (en 1840) puis Dirichlet (en 1850) utilisèrent les polygones de Voronoï dans leur étude des formes quadratiques. Voronoï généralisa leur travail à des dimensions supérieures en 1908. Quelques années plus tard, en 1934, Delaunay construisit une triangulation associée aux polygones de Voronoï et démontra la richesse de ses propriétés mathématiques.

Le package R *deldir* permet de calculer les polygones de Voronoï associés à un ensemble de points. La fonction `deldir` renvoie en sortie un objet représentable avec la fonction `plot`. Le package calcule également plusieurs statistiques associées aux polygones, telles que la surface de chaque polygone, ou encore le nombre de ses sommets (voir documentation détaillée :).

Il existe de nombreux algorithmes permettant de construire des polygones de Voronoï (le plus efficace est l'algorithme de Fortune (FORTUNE 1987). L'algorithme implémenté par la fonction `deldir` commence par construire une triangulation de Delaunay à partir des points de référence. Cette triangulation maximise l'angle minimal des triangles. Les sommets du diagramme de Voronoï sont les centres des cercles circonscrits des triangles de la triangulation de Delaunay. Les arêtes du diagramme de Voronoï sont sur les médiatrices des arêtes de la triangulation de Delaunay (l'algorithme est détaillé dans LEE et al. 1980).

Application avec R

```
#Packages nécessaires
library(deldir)
library(sp)

# Génération des points aléatoires
x <- rnorm(20, 0, 1.5)
y <- rnorm(20, 0, 1)

#Fonction "deldir" permettant de calculer les polygones de Voronoi
#à partir de deux jeux de coordonnées géographiques
vtess <- deldir(x, y)

#crée une fenêtre de travail
plot(x, y, type="n", asp=1)

#représente les points
```

```
points(x, y, pch=20,col ="red", cex=1)

#représente les polygones de Voronoi associés
plot(vtess, wlines="tess", wpoints="none", number=FALSE, add=TRUE, lty=1)
```

Pour passer d'une partition du territoire à un ensemble de points, on peut calculer les centroïdes des surfaces (figure 1.20).

Définition 1.3.2 — Centroïde d'une surface S. Point qui minimise la distance quadratique moyenne à tous les points de S :

$$\min_c \frac{1}{a(S)} \int_S ||x-c||^2 dx$$

$$c = \frac{1}{a(S)} \int_S x dx$$

Coordonnées de c : moyenne des coordonnées de **tous les points de S**

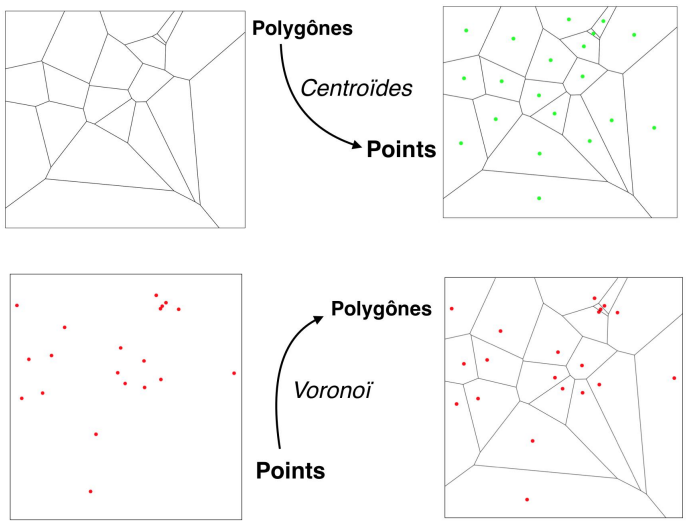


FIGURE 1.20 – Passage des points aux polygones et des polygones aux centroïdes

Application avec R

```
#Calcul des centroides des polygones
#A partir d'un fichier "Spatial Polygon Data Frame"

library(GISTools)
centroids<-  getSpPPolygonsLabptSlots(polygone)

plot(polygone)
points(centroids, pch = 20,col =  "Green", cex=0.5)
```

1.4 Exemples d'études utilisant des données spatiales agrégées

Le groupe européen pour l'intégration des données [3] souligne que représenter les données sur une carte avec une bonne résolution spatiale et temporelle permet de détecter des phénomènes invisibles autrement. Une représentation adaptée permet de bien comprendre la situation économique, sociale ou environnementale et de mettre en place des politiques publiques pertinentes. À travers les travaux réalisés par trois instituts de statistique européen, cette section illustre la variété des analyses descriptives utilisant des données spatiales : projet européen d'étude des taux de pauvreté régionaux ; analyse de la distance aux espaces verts par l'institut de statistique suédois ; analyse de la localisation optimale des éoliennes par l'institut de cartographie britannique.

1.4.1 Accès aux espaces verts - Statistique Suède

Augmenter l'accès aux espaces verts publics fait partie des objectifs environnementaux des politiques publiques suédoises. Dans de nombreuses municipalités suédoises, des débats opposent les partisans de la densification de l'habitat et ceux de la préservation des espaces verts.

L'association de données cartographiques satellitaires et d'informations statistiques localisées issues du recensement aide à mieux comprendre la situation sur le terrain, et ainsi à ajuster les politiques publiques. Cette étude rentre dans le cadre du 11ème objectif de développement durable des Nations Unies : "rendre les villes et les installations humaines inclusives, sûres, résistantes et durables".

En 2013, l'Institut de Statistique Suédois s'est appuyé sur l'analyse conjointe d'images satellites et de données administratives pour caractériser les espaces verts suédois, en fonction de leur statut de propriété, et de la qualité de leur végétation. La majorité des aires urbaines suédoises voient plus de 50 % de leur territoire recouvert par des espaces verts. En moyenne, trois quarts de ces espaces sont publics. Lidingö est la ville suédoise la plus recouverte d'espaces verts, puisque ceux-ci représentent environ 72 % de sa surface totale (figure 1.21). La deuxième partie de l'étude se concentre sur l'accessibilité à ces espaces verts. Grâce aux données du recensement de la population, l'Institut Suédois étudie la proportion d'adultes et d'enfants vivant à moins d'une certaine distance d'un espace vert public. Ainsi, **dans 26 aires urbaines suédoises, moins d'1 % de la population vit à plus de 300 mètres d'un espace vert accessible**. En revanche, dans certaines villes, comme Malmö, 15 % des enfants de moins de 6 ans n'ont pas accès à un espace vert à moins de 200 mètres.

1.4.2 Taux de pauvreté régionaux - programme européen ESPON

Le projet européen ESPON a pour objectif de favoriser l'harmonisation des politiques publiques européennes en mettant à disposition des décideurs des statistiques régionales pertinentes. Les différences de richesse entre les régions peuvent exacerber les sentiments d'exclusion et les tensions au niveau national. Cartographier le taux régional de pauvreté de la population permet de distinguer les zones les plus fragiles, et ainsi de mieux cibler les politiques d'aide au développement.

Le seuil de pauvreté (*At-Risk-of-Poverty (ARoP) threshold*) est défini comme 60 % de la médiane du niveau de vie national. Le seuil de pauvreté varie donc selon les pays (de 20 362 euros en Suisse à 5 520 en Grèce). Le taux de pauvreté (*ARoP rate*) est défini comme la part d'individus dont le niveau de vie est inférieur au seuil de pauvreté national. Dans la figure 1.22, on représente le ratio entre cet indicateur calculé au niveau infra-national (NUTS3) et le taux de pauvreté national. Cela permet **d'identifier les pays présentant les disparités régionales les plus importantes**, et de **visualiser au sein de chaque pays les zones les plus extrêmes**. Les plus

3. UN-GGIM : United Nations Commitee of Experts on Global Geospatial Information management - Working Group B - Europe

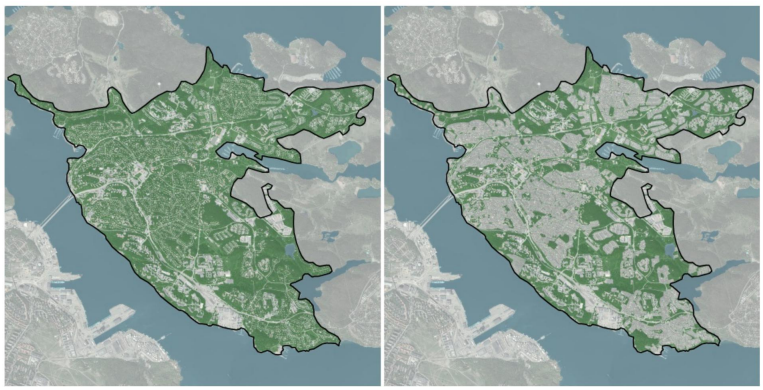


FIGURE 1.21 – Ville de Lidingö. Gauche : tous les espaces verts ; Droite : espaces verts accessibles au public (non privés)
Source : *Institut National de Statistiques Suédois*

grandes disparités inter-régionales dans les populations à risque sont observées en Turquie, Albanie, Hongrie, Allemagne, Croatie, Italie et Espagne. Les pays scandinaves, ainsi que les Pays-Bas, les États baltes, le Portugal et la Grèce ont une distribution plus uniforme des *ARoP rates*. Au sein des pays, on observe de faibles taux de pauvreté dans les alentours des capitales et des grandes villes, mais pas forcément au sein des villes elles-mêmes. Le taux de pauvreté est plus élevé dans les régions les moins accessibles, comme le sud de l'Italie, le centre de l'Espagne ou l'est de la Hongrie.

La représentation cartographique des taux de pauvreté ainsi définis permet d'aider la prise de décision publique, aussi bien au niveau national qu'européen. Dans ce but, le programme ESPON a mis en ligne de nombreuses analyses cartographiques de données démographiques et sociales : carte des ratios hommes-femmes par région, des différents profils d'innovation, des variations de taux d'emploi ou encore d'impact potentiel du changement climatique (https ://www.espon.eu/tools-maps).

1.4.3 Localisation optimale des éoliennes - Institut de cartographie de Grande Bretagne

Le gouvernement écossais a pour objectif d'augmenter la production d'énergie renouvelable d'ici 2020. Le conseil régional joue un rôle clé dans la préservation de l'équilibre local, en cherchant à développer le parc éolien tout en préservant la qualité de vie des habitants. Les données spatiales fournies par l'Institut de cartographie britannique (*Ordnance Survey*) ont beaucoup aidé à la prise de décision au niveau local.

Les objectifs de l'étude sont de donner des lignes directives claires et pratiques pour la localisation des champs d'éoliennes. L'étude doit prendre en compte de nombreux éléments environnementaux et sociaux comme les caractéristiques des paysages et le pittoresque des vues. Les données cartographiques doivent être suffisamment détaillées pour pouvoir servir de support aux planificateurs locaux, tout en restant suffisamment lisibles pour être comprises rapidement par l'ensemble des parties prenantes (figure 1.23).

Pour atteindre ces objectifs, l'*Ordnance Survey* a travaillé avec de nombreux experts locaux et utilisé de nombreuses bases géolocalisées. Les différents acteurs pouvaient suivre l'avancement de l'étude grâce à une cartographie interactive. Kevin Belton, ingénieur SIG et membre du conseil général régional souligne la valeur ajoutée de l'étude : "La communication d'informations complexes grâce à des données spatiales a permis au conseil régional de travailler avec un grand nombre de partenaires : aussi bien des membres du secteur public que des promoteurs commerciaux. Cette étude a

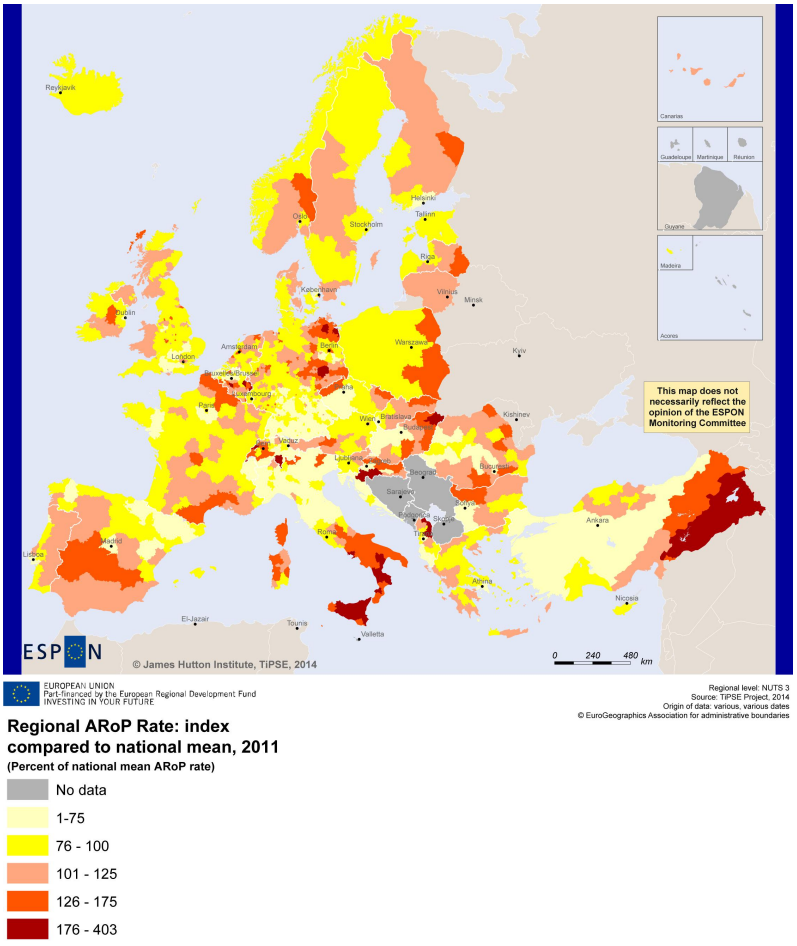


FIGURE 1.22 – Indicateur d'un risque élevé de pauvreté
Source : *Projet ESPON*
Note : pourcentage de l'indice régional rapporté à l'indice national

aidé les constructeurs à ne pas gaspiller de moyens dans des installations qui auraient été contraires aux politiques de préservation de l'environnement et de la qualité de vie des communautés locales."

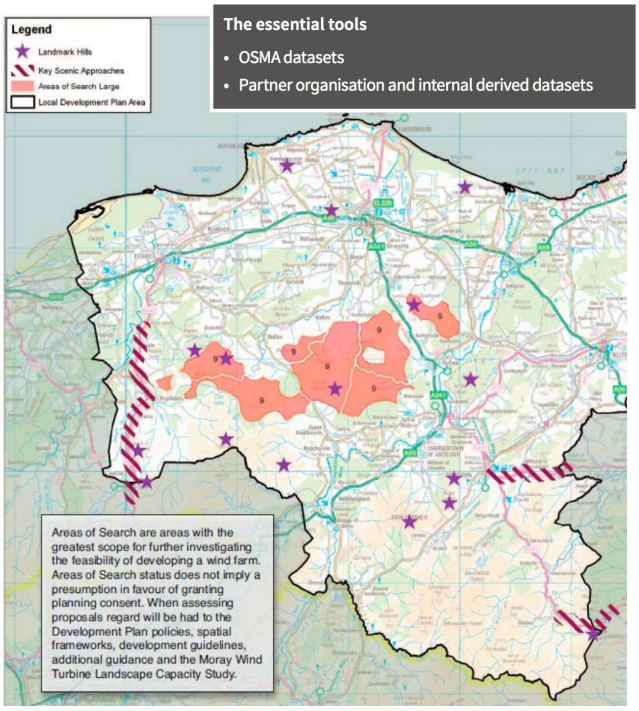


FIGURE 1.23 – Étude d'implantation d'éoliennes
Source : *Office National de Statistiques Britannique*
Note : Les aires de recherche "areas of search" sont les zones où il est le plus prometteur de lancer des investigations plus approfondies sur la possibilité d'y installer des éoliennes. Le statut d'"area of search" ne garantit pas l'octroi par les responsables de l'urbanisme local d'une autorisation de construction.

Références - Chapitre 1

AURENHAMMER, Franz (1991). « Voronoi diagrams : a survey of a fundamental geometric data structure ». *ACM Computing Surveys (CSUR)* 23.3, p. 345–405.

BIVAND, Roger S., Edzer PEBESMA et Virgilio GOMEZ-RUBIO (2008). *Applied spatial data analysis with R*. Springer.

CHILES, Jean-Paul et al. (2005). *Les pratiques de la géostatistique dans le domaine des sites et sols pollués*. GeoSiPol.

CRESSIE, Noel A.C. (1993b). « Statistics for spatial data : Wiley series in probability and statistics ». *Wiley-Interscience, New York* 15, p. 105–209.

FORTUNE, Steven (1987). « A sweepline algorithm for Voronoi diagrams ». *Algorithmica* 2.1-4, p. 153.

FOTHERINGHAM, A. Stewart et F. Benjamin ZHAN (1996). « A comparison of three exploratory methods for cluster detection in spatial point patterns ». *Geographical analysis* 28.3, p. 200–218.

GIVORD, Pauline et al. (2016). « Quels outils pour mesurer la ségrégation dans le système éducatif ? Une application à la composition sociale des collèges français ». *Education et formation*.

LEE, Der-Tsai et Bruce J SCHACHTER (1980). « Two algorithms for constructing a Delaunay triangulation ». *International Journal of Computer & Information Sciences* 9.3, p. 219–242.

2. Codifier la structure de voisinage

MARIE-PIERRE DE BELLEFON, VINCENT LOONIS, RONAN LE GLEUT
Insee

Résumé

Après avoir choisi l'échelle d'agrégation des données et effectué une première analyse descriptive grâce aux outils cartographiques, la deuxième étape d'une analyse spatiale est la définition du voisinage d'un objet. La définition du voisinage est indispensable pour mesurer la force des relations spatiales entre les objets, c'est-à-dire la façon dont les voisins s'influencent les uns les autres. Elle permet de calculer des indices d'autocorrélation spatiale, de mettre en œuvre les techniques d'économétrie spatiale, d'étudier la distribution spatiale des observations, mais aussi d'effectuer un échantillonnage spatial ou de partitionner un graphe.

L'enjeu de ce chapitre est de réussir à définir des relations de voisinage cohérentes avec les véritables interactions spatiales entre les objets. Ce chapitre présente plusieurs notions de voisinage, fondées sur la contiguïté ou sur les distances entre observations. La question du poids accordé à chaque voisin est aussi abordée. La mise en œuvre pratique s'appuie sur les packages R *spdep*, *tripack*, *spsurvey* et *tsp*.

R La lecture préalable du chapitre 1 : "Analyse spatiale descriptive" est recommandée.

2.1 Définir les voisins

2.1.1 Caractéristiques des relations entre objets spatiaux

Considérons une surface $\mathfrak{R}$. Cette surface peut être divisée en n zones mutuellement exclusives. Deux zones adjacentes sont séparées par une frontière commune. Les frontières peuvent naître de discontinuités spatiales (frontières administratives ou environnementales). Elles peuvent également être issues des polygones de Voronoï calculés à partir des points d'intérêt (voir chapitre 1 : "Analyse spatiale descriptive").

Encadré 2.1.1 — Définition mathématique des relations spatiales . Les relations spatiales $\mathscr{B}$ sont un sous-ensemble du produit cartésien $\mathbb{R}^2 \times \mathbb{R}^2 = \left\{(i,j) : i \in \mathbb{R}^2, j \in \mathbb{R}^2\right\}$ des couples (i,j) d'objets spatiaux, c'est-à-dire l'ensemble des couples (i,j) tels que i et j soient tous deux des objets spatiaux identifiés par leurs coordonnées géographiques, et que (i,j) soit différent de (j,i).
Un objet spatial ne peut pas être relié à lui-même : $(i,i) \notin \mathscr{B}$. De plus si $(i,j) \subseteq \mathscr{B}$ et $(j,i) \subseteq \mathscr{B}$ pour tout couple d'objets spatiaux, les relations spatiales sont dites *symétriques* (TIEFELSDORF 1998).

Les relations spatiales sont multidirectionnelles et multilatérales. Elles se distinguent en cela des relations temporelles qui n'autorisent que des relations séquentielles le long de l'axe passé-présent-futur.

La figure 2.1 illustre la démarche de codification des relations spatiales. Cette démarche permet de transcrire de manière systématique la complexité de l'espace géographique en un ensemble fini de données analysables par un ordinateur.
Tout d'abord, la zone d'étude est subdivisée en aires mutuellement exclusives. Chaque aire contient un point de référence (souvent son centroïde). Ensuite, les relations spatiales peuvent être spécifiées par un graphe de voisinage reliant les aires considérées comme voisines, ou par une matrice contenant les coordonnées géographiques des points de référence. La troisième étape consiste à coder le graphe dans une matrice de voisinage, ou à transformer les coordonnées géographiques en une matrice de distances.
La matrice de voisinage mesure la similarité entre les observations. Une valeur supérieure strictement à zéro indique que les observations sont considérées comme voisines. Par exemple, dans le cas de la matrice binaire présentée en figure 2.1 :

$$w_{ij} = \begin{cases} 1 & si \quad i \quad et \quad j \quad sont \quad reliés \quad dans \quad l'espace \\ 0 & sinon \end{cases} \tag{2.1}$$

Inversement, la matrice de distances mesure une dissimilarité entre zones. Plus d_{ij} est élevé, plus les zones diffèrent. Avec, si l'on utilise une distance euclidienne : $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$, α et β étant les coordonnées géographiques des observations.

La matrice de voisinage est utilisée dans l'étude des données spatiales surfaciques, tandis que la matrice de distances sert plutôt à la géostatistique (voir chapitre 5 : "Géostatistique"). On peut cependant passer de l'une à l'autre en définissant une distance minimale au-delà de laquelle les observations ne sont plus voisines.

La structure de la dépendance spatiale peut ne pas être géographique. Toute relation duale pertinente permet de définir un graphe de voisinage. Citons par exemple :

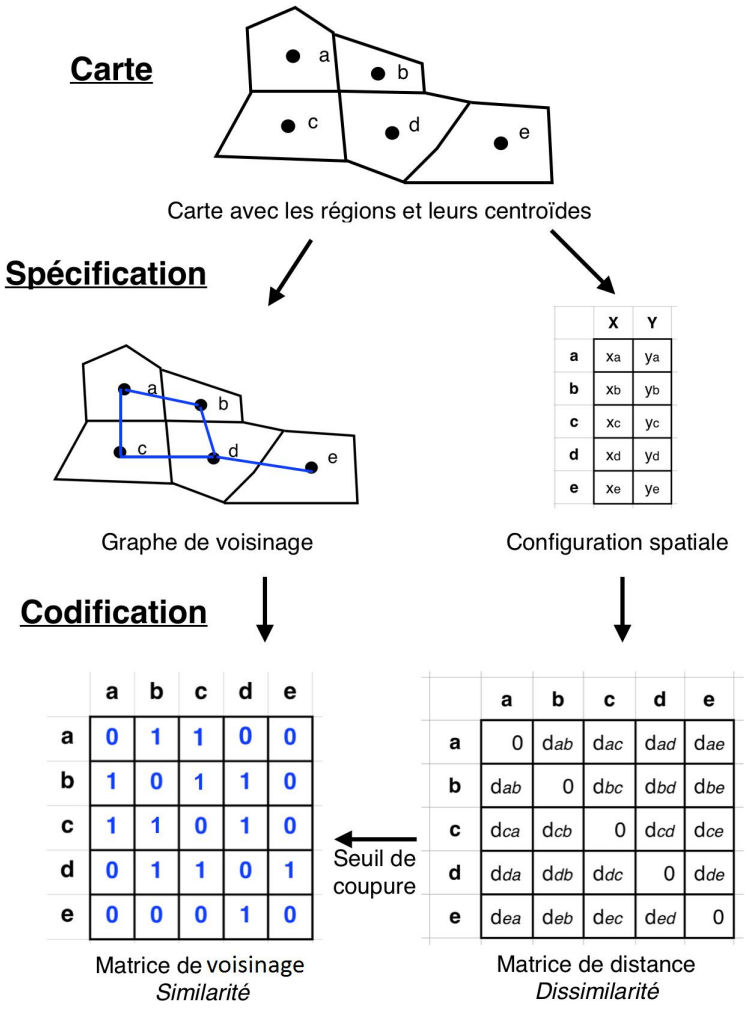


FIGURE 2.1 – Codification des relations spatiales
Source : TIEFELSDORF *1998*

— **au niveau des individus** : les liens d'amitié, la fréquence des communications, les citations dans les articles de recherche scientifique ;
— **au niveau des entreprises** : les liaisons siège-filiale, les similitudes en termes de marchés ;
— **au niveau international** : les alliances stratégiques, les flux commerciaux, l'appartenance commune à une organisation, les échanges culturels, les flux migratoires.

Les sections suivantes détaillent différentes spécifications de voisinage.

L'objet "liste de voisins" en R

Le package *spdep* permet de définir les relations de voisinage entre objets spatiaux. Dans R, la classe d'un objet définit l'ensemble de ses propriétés et la façon dont le statisticien peut l'utiliser. Les relations de voisinage sont enregistrées dans un objet de classe `nb`.
Soit n observations spatiales et *voisins_nb* l'objet spatial contenant les relations de voisinage associées. *voisins_nb* est une liste de longueur n. Chaque élément $[i]$ de la liste contient un vecteur avec l'index des voisins de l'élément d'index i. Si $[i]$ n'a pas de voisins, la liste contient uniquement 0. La liste contient également un vecteur de caractères associés à chaque zone de voisinage, ainsi qu'une valeur logique indiquant si la relation est symétrique ou pas (voir figure 2.2). Les informations principales sur l'objet *voisins_nb* peuvent être obtenues grâce à la fonction :

```
summary(voisins_nb)
```

La documentation du package *spdep* donne de plus amples informations (BIVAND et al. 2013b).

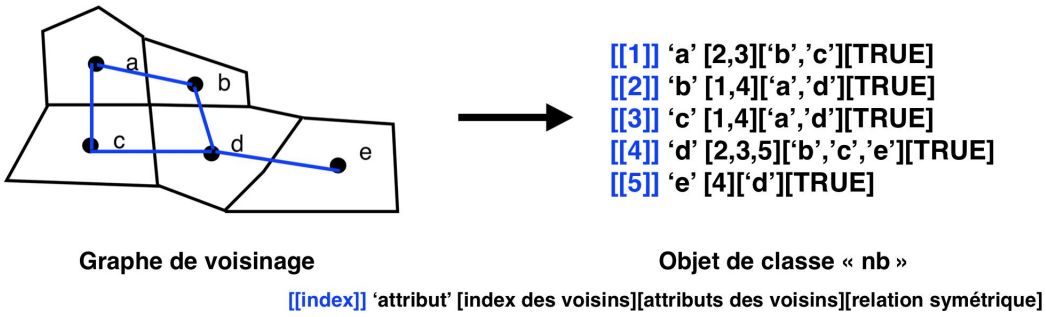


FIGURE 2.2 – La liste de voisins dans *spdep*

2.1.2 Définir les voisins en s'appuyant sur la distance

Dès lors qu'on dispose d'un ensemble de points répartis sur le territoire, on peut calculer les distances entre eux. Ces points peuvent être des lieux particuliers où l'information a été observée, ou l'ensemble des points représentatifs de chaque zone, par exemple leur centroïde. Dans ce cas, l'hypothèse sous-jacente est que la répartition de la valeur de la variable d'intérêt au sein de chaque zone est suffisamment homogène pour que l'approximation de l'attribuer à un unique point ne soit pas trop grossière.

Les graphes de voisinage matérialisent les liens entre les différentes entités. On les définit de façon à ce qu'ils représentent le plus fidèlement possible la structure spatiale sous-jacente. Il existe de nombreux graphes de voisinage différents. Nous présenterons ici les graphes fondés sur des notions géométriques et sur les voisins les plus proches.

Graphes de voisinage fondés sur des notions géométriques

La **triangulation de Delaunay** est une méthode géométrique qui relie les points sous forme de triangles tels que l'angle minimal de l'ensemble des triangles soit maximisé (cette triangulation

cherche à éviter les triangles "allongés"), voir figure 2.3 et 2.5a. La triangulation de Delaunay possède d'intéressantes propriétés géométriques et mathématiques. On peut cependant affiner la notion de voisinage.

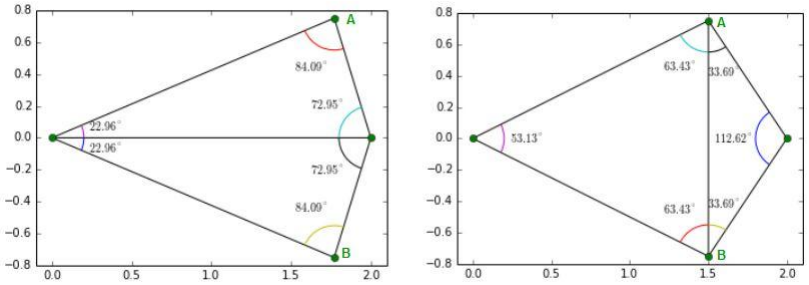


FIGURE 2.3 – Triangulation de Delaunay associée à différentes positions des points A et B
Source : *Gustavo [CC BY-SA 3.0 (https ://creativecommons.org/licenses/by-sa/3.0)], de Wikimedia Commons*

Le **graphe de la sphère d'influence** relie deux points si leurs "cercles du voisin le plus proche" se coupent. Le "cercle du voisin le plus proche" d'un point P est le plus grand cercle centré en P et qui ne contient pas d'autres points que P (voir figure 2.4 et 2.5b). Les graphes de la sphère d'influence ne sont pas nécessairement connectés, c'est-à-dire que tous les points de l'ensemble d'étude ne sont pas forcément reliés entre eux.

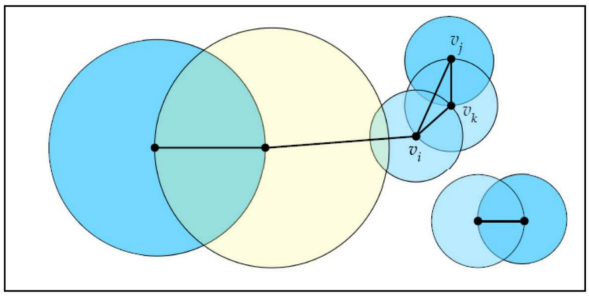


FIGURE 2.4 – Le graphe de la sphère d'influence d'un ensemble de points
Source : TOUSSAINT *2014*

Le **graphe de Gabriel** relie deux points p_i et p_j si et seulement si tous les autres points sont en dehors du cercle de diamètre $[p_i, p_j]$. Le graphe de Gabriel élimine certaine des liaisons du graphe de Delaunay, voir figure 2.5c.

Le **graphe des voisins relatifs** considère que deux points p_i et p_j sont voisins si

$$d(p_i, p_j) \leq max\,[d(p_i, p_k), d(p_j, p_k)] \quad \forall k = 1, ..., n \quad k \neq i, j \tag{2.2}$$

avec $d(p_i, p_j)$ la distance entre p_i et p_j. Le graphe des voisins relatifs impose moins de connexions que la triangulation de Delaunay ou le graphe de la sphère d'influence, voir figure 2.5d. TOUSSAINT 1980 juge qu'il s'adapte mieux aux données en imposant le moins de liaisons.

Les graphes de voisinage présentés ici sont tous des sous-graphes de la triangulation de Delaunay (voir figure 2.5). Ils ont l'avantage de ne laisser aucune unité sans voisins. En revanche, ils ne sont implémentés en R qu'avec la distance euclidienne, alors que d'autres types de distance, comme la distance du grand cercle, peuvent être plus adaptées à certaines études.

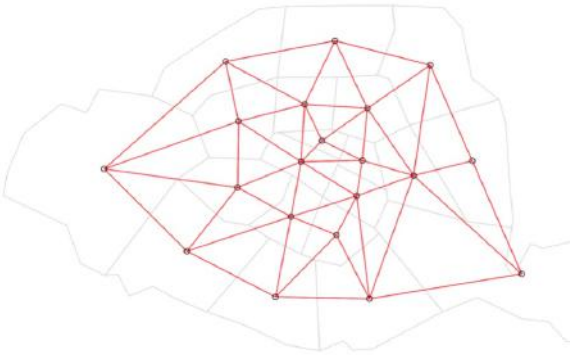
Application avec R

```
library(rgdal) #Pour importer les fichiers MIF/MID
library(maptools) #Pour importer les fichiers Shapefile
library(tripack) #Pour calculer les voisins basés sur la distance
library(spdep)

#Importation du fichier spatial
arr75 <- readOGR("~/ArmF.TAB", "ArmF")

#Voisins fondés sur la notion de graphe
#Le fichier en entrée est une matrice de coordonnées géographiques
#ou un objet de type SpatialPoints
coords <- coordinates(arr75)
IDs <- row.names(as(arr75,"data.frame"))

#Triangulation de Delaunay
Sy4_nb <- tri2nb(coords, row.names=IDs)
plot(arr75, border='lightgray')
plot(Sy4_nb,coordinates(arr75),add=TRUE,col='red')

#Graphe de la sphère d'influence
Sy5_nb <- graph2nb(soi.graph(Sy4_nb,coords),row.names=IDs)
plot(arr75, border='lightgray')
plot(Sy5_nb,coordinates(arr75),add=TRUE,col='red')

#Graphe de Gabriel
Sy6_nb <- graph2nb(gabrielneigh(coords), row.names=IDs)
plot(arr75, border='lightgray')
plot(Sy6_nb,coordinates(arr75),add=TRUE,col='red')

#Graphe des voisins relatifs
Sy7_nb <- graph2nb(relativeneigh(coords), row.names=IDs)
plot(arr75, border='lightgray')
plot(Sy7_nb,coordinates(arr75),add=TRUE,col='red')
```

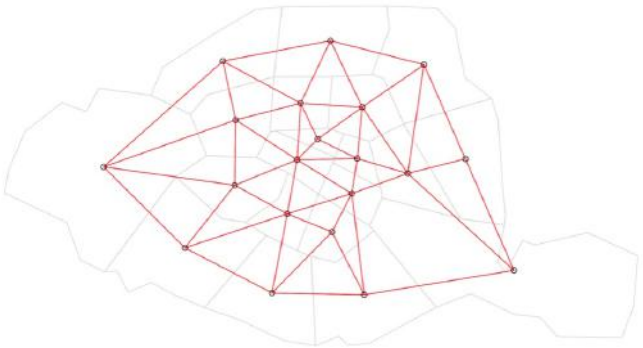
Graphes de voisinage fondés sur les voisins les plus proches

Une deuxième méthode consiste à sélectionner comme voisins les k points les plus proches (figure 2.6). Cette méthode a l'avantage de ne laisser aucun point sans voisin, ce qui n'est pas nécessaire pour conduire une analyse spatiale, mais reflète en général mieux la réalité (il est rare qu'une zone géographique soit complètement isolée). En revanche il est parfois difficile d'identifier la valeur k qui reflète les vraies relations spatiales sous-jacentes. Les graphes fondés sur les k voisins les plus proches ne sont pas nécessairement symétriques.

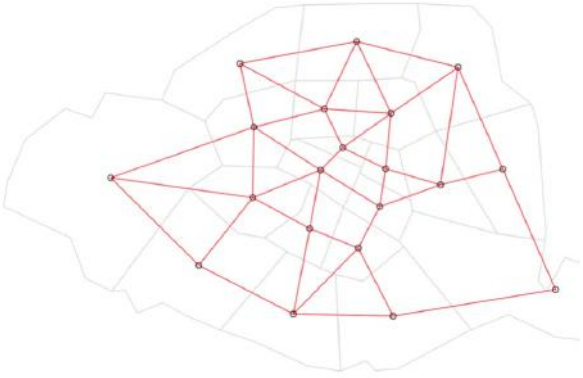
On peut également ne conserver que les points situés à une certaine distance. La fonction



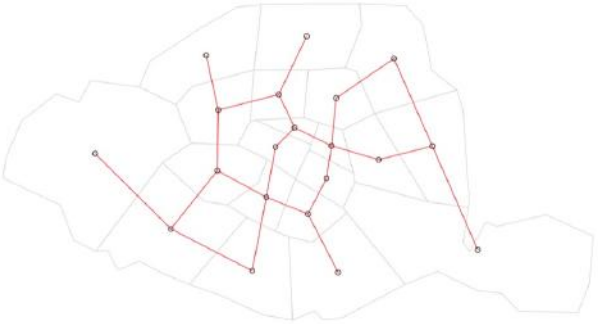
(a) Triangulation de Delaunay



(b) Graphe de la sphère d'influence



(c) Graphe de Gabriel



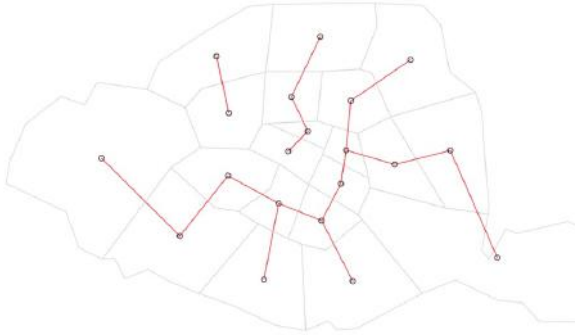
(d) Graphe des voisins relatifs

FIGURE 2.5 – Quatre graphes de voisinage des arrondissements parisiens fondés sur des notions géométriques

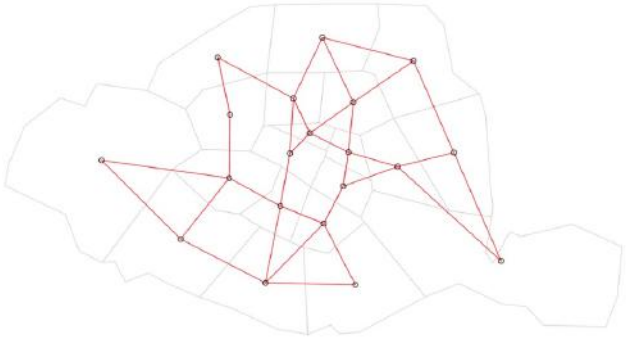
`nbdists` de R permet de calculer le vecteur des distances entre les voisins. On peut ainsi obtenir la distance minimale d_{min} au-delà de laquelle tous les points ont au moins un voisin, puis utiliser la fonction `dnearneighb` pour retenir comme voisins les seuls points situés entre les distances 0 et d_{min}. Cette méthode "de la distance minimale" n'est pas adaptée aux données irrégulièrement espacées car la distance minimale nécessaire pour qu'un point relativement isolé ait au moins un voisin est beaucoup plus élevée que la distance du plus proche voisin d'un point situé dans une zone dense. Il y aura donc de grandes disparités dans le nombre de voisins (BIVAND et al. 2013b), voir figure 2.6d.

Application avec R - Source : BIVAND *et al. 2013b*

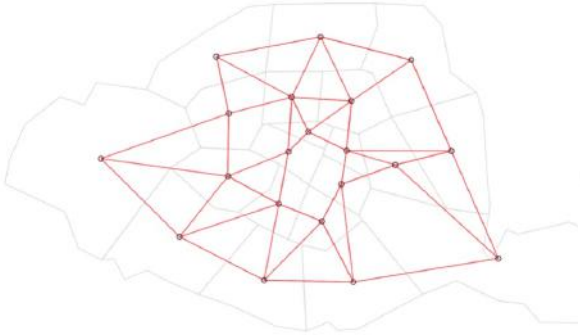
```
#Graphes fondés sur les plus proches voisins
Sy8_nb <- knn2nb(knearneigh(coords,k=1),row.names=IDs)
Sy9_nb <- knn2nb(knearneigh(coords,k=2),row.names=IDs)
Sy10_nb <- knn2nb(knearneigh(coords,k=3),row.names=IDs)

plot(arr75, border='lightgray')
plot(Sy8_nb,coordinates(arr75),add=TRUE,col='red')

#Etude de la distance moyenne du voisin le plus proche
dsts <- unlist(nbdists(Sy8_nb,coords))
summary(dsts)
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    820    1188    1678    1707    2016    3412
max_1nn <- max(dsts)

#Calcul et représentation des voisins à la distance minimale
Sy11_nb <- dnearneigh(coords, d1=0, d2=max_1nn, row.names=IDs)
plot(arr75, border='lightgray')
plot(Sy11_nb,coordinates(arr75),add=TRUE,col='red')
```



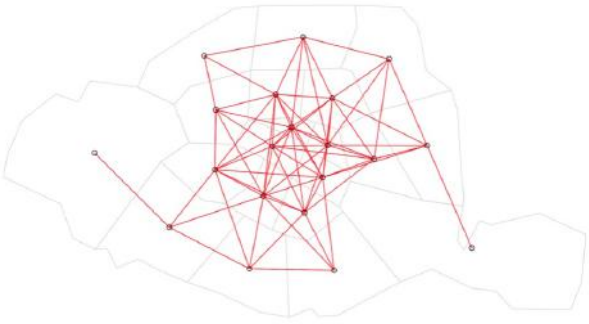
(a) Voisin le plus proche



(b) Deux plus proches voisins



(c) Trois plus proches voisins



(d) Voisins à la distance minimale

FIGURE 2.6 – Quatre graphes fondés sur les plus proches voisins des arrondissements parisiens

2.1.3 Définir les voisins en s'appuyant sur la contiguïté

Lorsque les données surfaciques consistent en une partition de l'ensemble du territoire, la notion de "distance entre les observations" peut devenir assez ambiguë. L'exemple 2.1 illustre les limites de l'utilisation de la distance entre centroïdes pour définir la notion de voisinage.

■ **Exemple 2.1 — Ambiguïté de la notion de distance entre centroïdes.** Soient R_1, R_2, R_3 trois zones distinctes. On peut considérer que comme R_2 et R_3 sont séparées dans l'espace, mais toutes les deux adjacentes à R_1, elles sont toutes les deux plus proches de R_1 que l'une de l'autre. Cependant, les centroïdes de ces zones sont équidistants entre eux (voir figure 2.7). Résumer la proximité entre zones par la distance entre centroïdes conduit à perdre une partie de la richesse des relations spatiales.

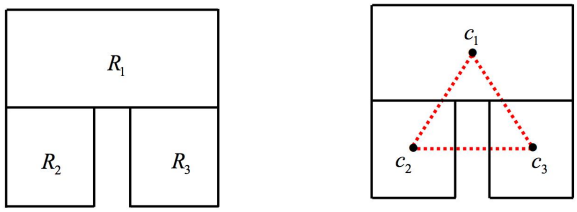


FIGURE 2.7 – Gauche : trois zones - Droite : distances entre centroïdes
Source : SMITH *2016*

■

Cette sous-section introduit différentes notions de contiguïté et présente la façon dont le package R *spdep* permet de créer une liste de voisins.

Au sens de la contiguïté **Rook**, les voisins possèdent au moins un segment de frontière commune. Cela correspond aux déplacements de la Tour du jeu d'échecs. Pour que deux zones soient voisines au sens de la contiguïté **Queen**, il suffit qu'elles partagent un point de frontière commune. Cela correspond aux déplacements de la Reine du jeu d'échecs. La figure 2.8 illustre ces notions dans le cas d'une grille régulière de points. Quand les polygones ont une forme et une surface irrégulières, les différences entre voisinage Rook et Queen deviennent plus difficiles à appréhender. Notons également qu'une zone très étendue entourée de plus petites zones aura un nombre de voisins beaucoup plus important que ses voisines.
Le voisinage au sens de la contiguïté est souvent utilisé pour étudier des données démographiques et sociales, pour lesquelles être d'un côté ou de l'autre d'une frontière administrative peut avoir plus d'importance qu'être situé à une certaine distance l'une de l'autre.

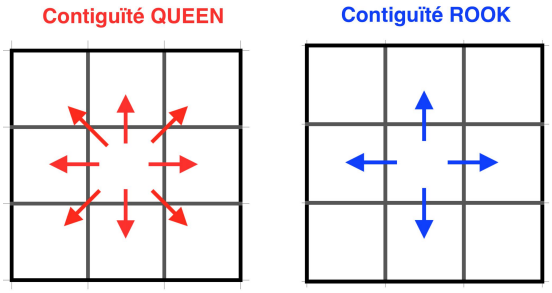


FIGURE 2.8 – Définition de la contiguïté Queen et Rook

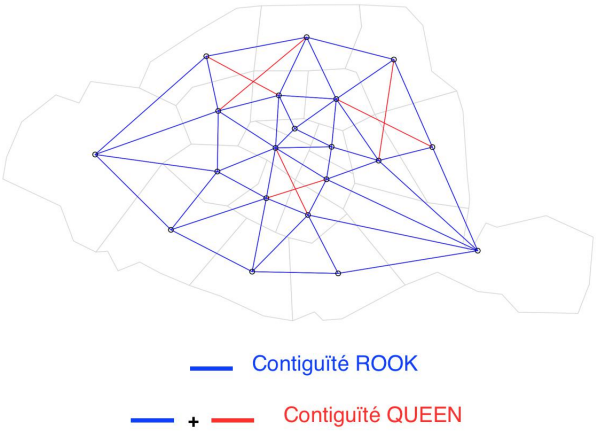


FIGURE 2.9 – Contiguïté Queen et Rook des arrondissements parisiens

Application avec R

Construction des graphes de voisinage Queen et Rook pour les arrondissements parisiens (figure 2.9)

```
#Le fichier en entrée est un fichier SpatialPolygons
#Extraction de la liste des voisins au sens QUEEN (par défaut)
arr75.nb<-  poly2nb(arr75)
```

```
#Extraction de la liste des voisins au sens ROOK
arr75.nb.ROOK <- poly2nb(arr75, queen=FALSE)

#Représentation visuelle des voisins :
plot(arr75, border='lightgray')
plot(arr75.nb, coordinates(arr75),add=TRUE,col='red')
plot(arr75.nb.ROOK, coordinates(arr75),add=TRUE,col='blue')
```

2.1.4 Définir les voisins en s'appuyant sur l'optimisation d'une trajectoire

Autour du voyageur de commerce

Certaines méthodes comme celle de l'échantillonnage spatial (voir chapitre 10 : "Échantillonnage spatial") nécessitent un tri préalable des données. Quand ces dernières sont caractérisées par deux variables (à savoir leurs coordonnées géographiques dans le plan), le choix de la méthode de tri est un problème théorique complexe.

Une solution consiste à faire passer un *chemin* par l'ensemble des points puis à les trier selon leur ordre d'apparition quand on parcourt le chemin. Les voisins d'un point donné sont alors les points situés juste avant ou juste après sur le chemin.

Parmi l'ensemble des chemins possibles, certains ont des caractéristiques plus adaptées à l'objectif recherché, comme par exemple celui de réduire la variance d'échantillonnage. C'est le cas du *plus court chemin*. Il minimise la somme des distances entre deux points consécutifs. Ce chemin qui ne fixe pas de contraintes particulières sur le point de départ ou d'arrivée est connu dans la littérature de la théorie des graphes comme *chemin de Hamilton* (figure 2.11b) associé à un graphe dont les arêtes sont pondérées. Un cas particulier célèbre de *plus court chemin* est celui dit du voyageur de commerce. Il représente le chemin que doit suivre un voyageur de commerce pour visiter l'ensemble de ses clients, tout en minimisant la distance parcourue et en rentrant chez lui le soir. Un tel chemin correspond à un cycle hamiltonien (figure 2.11c).

La recherche d'un plus court chemin est un problème classique d'optimisation dans le cadre de la théorie des graphes. Il est apparu notamment pour la résolution par Euler du problème des sept ponts de Königsberg [1]. Il intervient dans les questions relatives aux graphes eulériens ou hamiltoniens [2]. Il n'existe pas aujourd'hui d'algorithme en temps polynomial permettant de trouver le plus court chemin. Quand le nombre de points est grand, la recherche du chemin optimal passe par des heuristiques [3] conduisant à un optimum local. Elles sont disponibles dans le package *TSP* de R (HAHSLER et al. 2017).

Quand la distance est euclidienne et le nombre de points raisonnable, de l'ordre de quelques centaines, une solution exacte peut être trouvée grâce au programme `concorde` (APPLEGATE et al. 2006). Ce programme peut être appelé directement depuis R et le package *TSP*.

Enfin, la recherche d'un chemin hamiltonien à partir d'une matrice de distances est équivalente à celle d'un cycle hamiltonien pour peu que l'on rajoute une ligne et une colonne formées de 0 à la matrice originelle (GARFINKEL 1985). Le package *TSP* prévoit explicitement ce cas avec la fonction `insert-dummy`.

1. La question étudiée par Euler était : dans la ville de Königsberg peut-on faire une promenade en parcourant chacun des 7 ponts une fois et une seule ?

2. Un graphe eulérien est un graphe que l'on peut parcourir en partant d'un sommet quelconque et en empruntant exactement une fois chaque arête pour revenir au sommet de départ. Il correspond à un dessin qu'on peut tracer sans lever le crayon. Un graphe hamiltonien est un graphe que l'on peut parcourir en passant par tous les sommets une fois et une seule. Un graphe hamiltonien n'est pas nécessairement eulérien car dans un cycle hamiltonien, on peut très bien négliger de passer par certaines arêtes.

3. une heuristique est une méthode de calcul qui fournit rapidement (en temps polynomial) une solution réalisable, pas nécessairement optimale.

D'autres méthodes

La méthode *general randomized tessellation stratified* (GRTS , STEVENS JR et al. 2004) est populaire en échantillonnage spatial, puisqu'elle permet d'obtenir un échantillon spatialement réparti pour une population finie d'individus (unités distinctes et identifiables de dimension 0 d'une population discrète, par exemple des arbres d'une forêt), une population linéaire (unités continues de dimension 1, e.g. des rivières) ou une population de surfaces (unités continues de dimension 2, par exemple des forêts). Elle s'appuie sur un chemin construit à partir d'une classe de fonctions appelée *quadrant-recursive* (MARK 1990), permettant d'assurer que certaines relations de proximité de l'espace à deux dimensions soient toujours préservées dans l'espace à une dimension.

L'idée de la méthode est de projeter les coordonnées sur un carré unitaire, puis de découper ce carré en quatre cellules, chacune d'entre elles étant à nouveau divisée en quatre sous-cellules, etc. À chaque cellule on attribue une valeur résultant de l'ordre dans lequel le découpage a été effectué, ce qui permet finalement de placer les unités sur le chemin parcourant l'espace à deux dimensions.

La figure 2.10 montre les premières étapes du découpage, qui peut être mis en œuvre avec le package `spsurvey` de R (KINCAID et al. 2016). La méthode GRTS conduit cependant à créer de *grands sauts* (figures 2.11d) dans les chemins, ce qui peut affecter la précision des estimations.

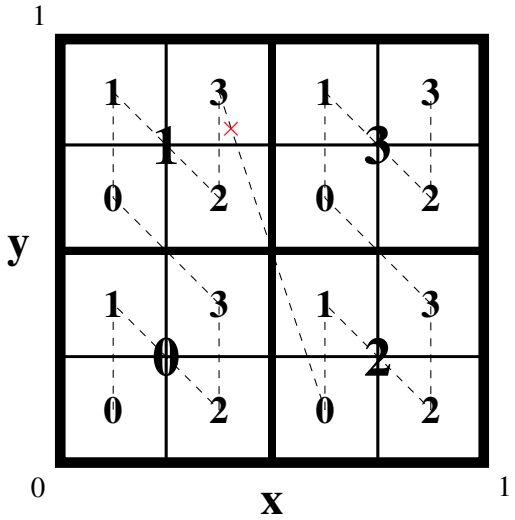
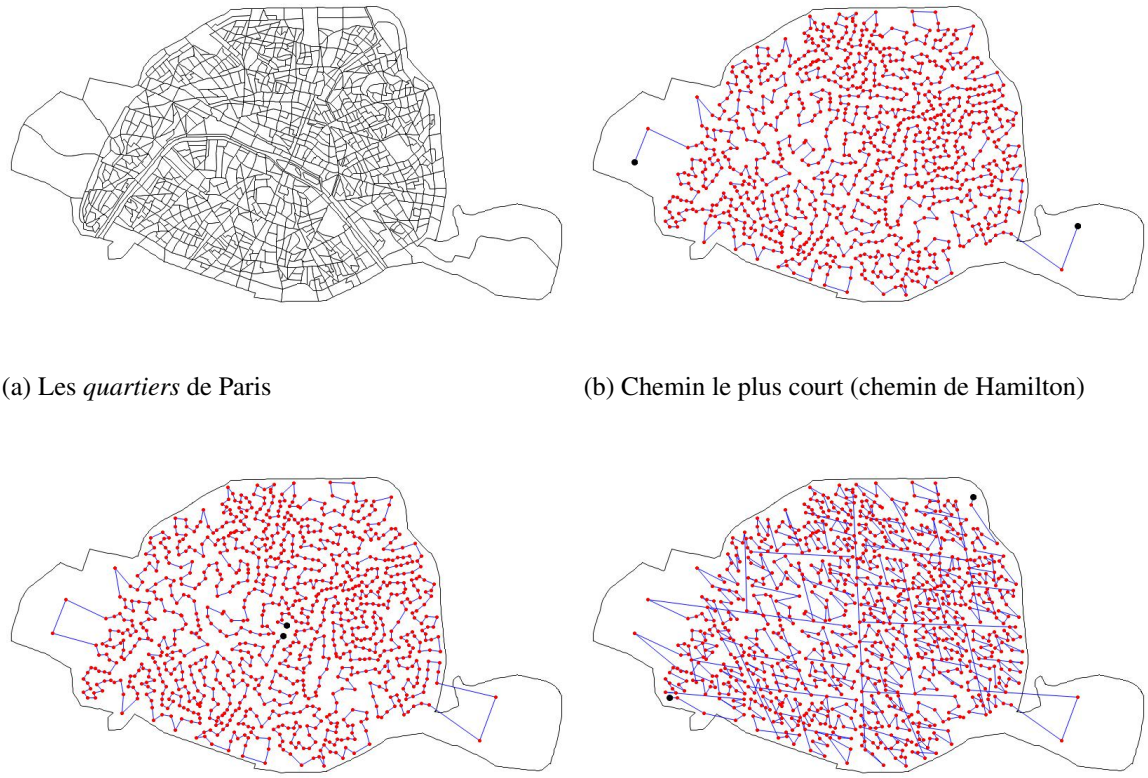


FIGURE 2.10 – Construction d'un chemin avec la méthode GRTS
Note : À l'unité dont la position est une croix rouge est associée la valeur "13" permettant ensuite de la positionner sur le chemin.

Application avec R - Source : Recherche d'un plus court chemin

```
library(TSP)
library(miscTools)

#Il faut télécharger l'utilitaire concorde à cette adresse :

http://www.tsp.gatech.edu/concorde/downloads/downloads.htm
#et l'appeler depuis R

Sys.setenv(PATH=paste(Sys.getenv("PATH"),"z:/cygwin/App/Runtime/Cygwin/bin"
    ,sep=";"))
concorde_path("Z:/concorde/")
```

```
#Les données en entrée sont une matrice de distances
test <-as.matrix(read.csv("U:/paris.csv",header=FALSE,sep="\t"))

#les erreurs d'arrondis peuvent conduire à ce que la matrice ne soit
    parfaitement symétrique.

tsp <-(symMatrix(test[upper.tri(test, TRUE)],nrow =nrow(test), byrow=TRUE))
#on crée un objet lisible par TSP
tsp<-TSP(tsp)
#On applique la méthode concorde à cet objet.
tour<-solve_TSP(tsp, method = "concorde")
```



(a) Les *quartiers* de Paris

(b) Chemin le plus court (chemin de Hamilton)

(c) Cycle hamiltonien, chemin du voyageur de commerce

(d) Construction d'un chemin avec la méthode GRTS

FIGURE 2.11 – Recherche de chemins passant par tous les *quartiers* de Paris

2.2 Accorder des poids aux voisins

2.2.1 Passer d'une liste de voisins à une matrice de poids

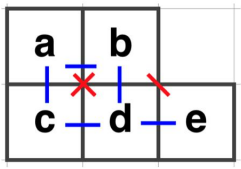
Une fois le graphe de voisinage défini et codifié sous forme d'une liste de voisins, on transforme la liaison entre les points i et j en l'élément w_{ij} de la matrice de poids $\mathbf{W}$. La matrice de poids $\mathbf{W}$

est "l'expression formelle de la dépendance spatiale entre observations" (ANSELIN et al. 1988).

Définition de la matrice de poids

— Le plus couramment, la matrice de poids est une matrice de contiguïté binaire (voir figure 2.12) :

$$w_{ij} = \begin{cases} 1 & si \quad i \quad et \quad j \quad sont \quad reliés \quad dans \quad l'espace \\ 0 & sinon. \end{cases} \tag{2.3}$$



Contiguïté ROOK

	a	b	c	d	e	Somme des poids des voisins
a	0	1	1	0	0	2
b	1	0	0	1	0	2
c	1	0	0	1	0	2
d	0	1	1	0	1	3
e	0	0	0	1	0	1

Contiguïté QUEEN

	a	b	c	d	e	Somme des poids des voisins
a	0	1	1	1	0	3
b	1	0	1	1	1	4
c	1	1	0	1	0	3
d	1	1	1	0	1	4
e	0	1	0	1	0	2

FIGURE 2.12 – Matrice de poids binaire

— Les matrices de poids peuvent également tenir compte de la distance entre les zones géographiques, les relations devenant plus faibles avec la distance : 1 si $d < d_0$, $\frac{1}{d^\alpha}$, ou $e^{-\alpha d}$ avec α un paramètre estimé ou défini *a priori*. Utiliser une distance maximale au-delà de laquelle $w_{ij} = 0$ permet de limiter le nombre d'éléments non nuls. Comme décrit en 2.1.2, lorsque les tailles des zones sont hétérogènes, on risque alors une grande variabilité du nombre de voisins.

— Enfin, certaines matrices tiennent compte de la force des relations entre les zones. Le poids peut par exemple être défini par $\frac{b_{ij}^\alpha}{d_{ij}^\beta}$ avec b_{ij} une mesure de la force des relations entre les zones i et j (qui n'est pas forcément symétrique), telle que le pourcentage de frontières communes, la population totale, la richesse et d_{ij} la distance entre les zones.

Certains travaux économétriques cherchent à endogénéiser les matrices de poids, mais elles sont considérées comme exogènes dans la plupart des applications d'économétrie spatiale (ANSELIN 2013). En général, les poids de voisinage ne doivent donc pas être des fonctions du phénomène qu'on cherche à expliquer.

L'objet "liste de poids" en R

La fonction `nb2listw` du package `spdep` permet de transformer un objet "liste de voisins" en un objet "liste de poids". Il est important de noter que l'objet "liste de poids", qui correspond à la matrice de poids décrite précédemment n'est pas une matrice $n \times n$ telle qu'on la représente théoriquement. Il s'agit d'une liste contenant le style de normalisation, puis pour chaque observation : son attribut, la liste des numéros d'observation de ses voisins, la liste des attributs de ses voisins et la liste des

poids de ses voisins. On parle souvent de *sparse matrix* ou *matrices creuses*.
Lorsqu'une zone n'a pas de voisins, l'option `zero.policy=TRUE` permet de générer tout de même une liste de poids, qui prend la valeur 'zéro' pour les observations sans voisins. (Si l'option est `FALSE`, un message d'erreur est généré).

Application avec R

```
#Matrice basée sur la contiguïté
#La fonction nb2listw convertit tout objet de type "nb" en une #liste de
    poids
arr75.lw <- nb2listw(arr75.nb)

#Matrice fondée sur la distance
#La fonction mat2listw convertit une matrice en une liste de poids
library(fields) #pour calculer la distance entre deux points
coords <- coordinates(arr75)
distance <- rdist(coords,coords)
diag(distance) <- 0
distance[distance >=100000] <- 0
#le poids décroit comme le carré de la distance, dans un rayon de 100km
dist <- 1.e12 %/% (distance*distance)
dist[dist >=1.e15] <- 0
dist.w <- mat2listw(dist,row.names=NULL)
```

Normalisation de la matrice de poids

La somme des poids des voisins d'une zone est appelée son *degré de liaison*. Si on ne normalise pas la matrice de poids (*schéma de codage "B"*), ce degré de liaison dépend du nombre de ses voisins, ce qui crée une hétérogénéité entre les zones. À la suite de TIEFELSDORF 1998, on peut distinguer quatre types de normalisation :

— Normalisation en ligne (*schéma de codage "W"*) : pour une zone, le poids accordé à chaque voisin est divisé par la somme des poids de ses voisins : $\sum_{j=1}^{n} w_{ij} = 1$. Cette standardisation facilite l'interprétation de la matrice de poids, puisque $\sum_{j=1}^{n} w_{ij}x_j$ représente la moyenne de la variable x sur tous les voisins de l'observation i. Chaque poids w_{ij} peut être interprété comme la fraction de l'influence spatiale subie par l'observation i imputable à j. En revanche, cette normalisation implique une certaine compétition entre les voisins : moins une zone a de voisins, plus ceux-ci ont un poids important. De plus, quand les poids sont inversement proportionnels à la distance entre les zones, normaliser en ligne rend difficile leur interprétation.

— Normalisation globale (*schéma de codage "C"*) : les poids sont standardisés de sorte que la somme de tous les poids soit égale au nombre total d'entités : tous les poids sont multipliés par $\frac{n}{\sum_{j=1}^{n}\sum_{i=1}^{n} w_{ij}}$.

— Normalisation uniforme (*schéma de codage "U"*) : les poids sont standardisés de sorte que la somme de tous les poids soit égale à 1 : $\sum_{j=1}^{n}\sum_{i=1}^{n} w_{ij} = 1$.

— Normalisation par stabilisation de la variance (*schéma de codage "S"*) : soit $\mathbf{q}$ le vecteur défini par : $\mathbf{q} = (\sqrt{\sum_{j=1}^{n} w_{1j}^2}, \sqrt{\sum_{j=1}^{n} w_{2j}^2}, ..., \sqrt{\sum_{j=1}^{n} w_{nj}^2})^T$.

Soit la matrice $\mathbf{S}^* = [diag(\mathbf{q})]^{-1}\mathbf{W}$.[4] À partir de $\mathbf{S}^*$, on calcule $Q = \sum_{j=1}^{n}\sum_{i=1}^{n} s_{ij}^*$ puis on en déduit la matrice de poids normalisée : $\mathbf{S} = \frac{n}{Q}\mathbf{S}^*$.

La normalisation par stabilisation de la variance a été introduite par Tiefelsdorf afin de réduire l'hétérogénéité dans les poids liée aux différences de taille et de nombre de voisins entre les zones. En effet, la normalisation en ligne donne plus de poids aux observations situées en bordure de la zone d'étude, avec un faible nombre de voisins. Au contraire, avec une normalisation globale ou uniforme, les observations situées au centre de la zone d'étude, avec un grand nombre de voisins, sont soumises à plus d'influences extérieures que les zones frontalières. Cette hétérogénéité peut avoir un impact significatif sur les résultats des tests d'autocorrélation spatiale.

Les poids de la matrice normalisée suivant le schéma "S" varient moins que ceux de la matrice normalisée suivant le schéma "W". La somme des poids des lignes varie plus pour le style "S" que pour le style "W", mais moins que pour les styles "B", "C" et "U" (BIVAND et al. 2013b).

Que le schéma de codage soit en ligne, global ou par stabilisation de la variance, la somme totale des éléments de la matrice vaut toujours n, ce qui permet aux statistiques d'autocorrélation spatiale utilisant la matrice d'être comparables entre elles.

Application avec R

```
#L'option style permet de définir le type de normalisation
arr75.lw <- nb2listw(arr75.nb,zero.policy=TRUE, style="W")
names(arr75.lw)
## [1] "style"      "neighbours" "weights"
summary(unlist(arr75.lw$weights))
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.1250  0.1667  0.1833  0.1961  0.2500  0.3333
```

2.2.2 Importance du choix de la matrice de poids

Lorsqu'on cherche à tester l'importance des relations économiques ou sociales entre certaines variables, la localisation géographique des observations est un paramètre clé. D'une part, les observations situées dans la même zone géographique sont soumises aux mêmes paramètres extérieurs (climat, pollution, etc.) ; d'autre part les observations voisines s'influencent mutuellement. Les modèles d'économétrie spatiale prennent en compte ces différentes interactions. Ces modèles utilisent la spécification du voisinage, par l'intermédiaire de la matrice de poids $\mathbf{W}$. Au sein de la communauté scientifique, les avis diffèrent sur l'influence de la définition de la matrice de poids sur les résultats.

BHATTACHARJEE et al. 2005 remarquent que : "Le choix des poids est souvent arbitraire [...] et le résultat des études varie considérablement en fonction de la définition des poids spatiaux". Une mauvaise spécification de $\mathbf{W}$ conduirait à de fausses conclusions. Ceci dit, puisque différentes méthodes de construction de la matrice de poids sont envisageables, "[...] il est possible que l'une des méthodes mène à des résultats pertinents, mais le risque d'une mauvaise spécification pèsera toujours sur le modèle choisi" (GETIS et al. 2004).

4. $diag(\mathbf{q})$ est une matrice diagonale avec les composantes de $\mathbf{q}$ sur sa diagonale principale

L'objectif est que les poids w_{ij} reflètent le plus fidèlement possible les interactions entre observations. Les hypothèses sous-jacentes peuvent s'appuyer sur des modèles économiques ou sociologiques. Par exemple, des poids nuls au-delà d'une certaine distance seront justifiés par le fait que l'influence d'une zone d'emploi sur son environnement est contrainte par la mobilité des individus, elle-même limitée par leur temps de trajet. HARRIS et al. 2011 soulignent cependant que le concept de 'distance' est lui-même flou. La distance est souvent définie par une distance géométrique entre deux points représentatifs des zones d'étude. Mais la distance peut également être un temps de transport entre deux régions (temps minimal, ou temps en empruntant la route la moins onéreuse), ou encore être proportionnelle aux échanges entre les zones. Pour HARRIS et al. 2011, "la conséquence de l'utilisation de mesures liées à la contiguïté ou à la distance pour pondérer les observations des régions voisines est qu'on impose une structure d'interaction spatiale dont on est incapable de tester la fiabilité, et qui est potentiellement mal spécifiée."

HARRIS et al. 2011 présentent quelques approches alternatives de construction de la matrice de poids. Ces méthodes ont pour objectif de diminuer au maximum les hypothèses *ad hoc* dans la spécification des matrices. Cependant aucune méthode n'arrive à s'en défaire totalement.

Tous les chercheurs ne sont pas aussi pessimistes : LESAGE et al. 2010 considèrent que la croyance selon laquelle la définition de la matrice de poids a une influence cruciale sur les résultats est due à des erreurs d'interprétation des coefficients des modèles d'économétrie spatiale, ou à des erreurs dans la spécification des modèles. Cette croyance serait, selon eux : "le plus gros mythe de l'économétrie spatiale". Ils soutiennent que si l'on s'intéresse à l'effet moyen des variables explicatives sur les variables dépendantes, les différences de spécification de la matrice de poids n'ont pas d'influence significative sur les résultats. LESAGE et al. 2010 reconnaissent cependant qu'il reste encore beaucoup à faire pour mieux caractériser la notion d'équivalence entre matrices.

Références - Chapitre 2

ANSELIN, Luc (2013). *Spatial econometrics : methods and models*. T. 4. Springer Science & Business Media.

ANSELIN, Luc et Daniel A GRIFFITH (1988). « Do spatial effects really matter in regression analysis ? » *Papers in Regional Science* 65.1, p. 11–34.

APPLEGATE, David et al. (2006). *Concorde TSP solver*.

BHATTACHARJEE, Arnab et Chris JENSEN-BUTLER (2005). « Estimation of spatial weights matrix in a spatial error model, with an application to diffusion in housing demand ». CRIEFF Discussion Papers.

BIVAND, Roger S, Edzer PEBESMA et Virgilio GOMEZ-RUBIO (2013b). « Spatial Neighbors ». *Applied Spatial Data Analysis with R*. Springer, p. 83–125.

GARFINKEL, R.S. (1985). « Motivation and modelling (chapter 2) ». *E. L. Lawler, J. K. Lenstra, A.H.G. Rinnooy Kan, D. B. Shmoys (eds.) The traveling salesman problem - A guided tour of combinatorial optimization,* Wiley & Sons.

GETIS, A et J ALDSTADT (2004). « On the specification of the spatial weights matrix ». *Geographical Analysis* 35.

HAHSLER, Michael et Kurt HORNIK (2017). *TSP : Traveling Salesperson Problem (TSP)*. R package version 1.1-5. URL : `https://CRAN.R-project.org/package=TSP`.

HARRIS, Richard, John MOFFAT et Victoria KRAVTSOVA (2011). « In search of 'W' ». *Spatial Economic Analysis* 6.3, p. 249–270.

KINCAID, Thomas M. et Anthony R. OLSEN (2016). *spsurvey : Spatial Survey Design and Analysis*. R package version 3.3.

LESAGE, James P et R Kelley PACE (2010). « The biggest myth in spatial econometrics ». *Available at SSRN 1725503*.

MARK, David M (1990). « Neighbor-based properties of some orderings of two-dimensional space ». *Geographical Analysis* 22.2, p. 145–157.

SMITH, Tony E. (2016). *Notebook on Spatial Data Analysis*. http ://www.seas.upenn.edu/ ese502/notebook.

STEVENS JR, Don L et Anthony R OLSEN (2004). « Spatially balanced sampling of natural resources ». *Journal of the American Statistical Association* 99.465, p. 262–278.

TIEFELSDORF, Michael (1998). « Modelling spatial processes : The identification and analysis of spatial relationships in regression residuals by means of Moran's I (Germany) ». Thèse de doct. Université Wilfrid Laurier.

TOUSSAINT, Godfried T (1980). « The relative neighbourhood graph of a finite planar set ». *Pattern recognition* 12.4, p. 261–268.

— (2014). « The sphere of influence graph : Theory and applications ». *International Journal of Information Theory and Computer Science* 14.2.

II Partie 2 : Mesurer l'importance des effets spatiaux

3. Indices d'autocorrélation spatiale

BOUAYAD AGHA SALIMA
GAINS (TEPP) et Crest
Le Mans Université

MARIE-PIERRE DE BELLEFON
Insee

Résumé

Les indices d'autocorrélation spatiale permettent de mesurer la dépendance spatiale entre les valeurs d'une même variable en différents endroits de l'espace. Plus les valeurs des observations sont influencées par les valeurs des observations qui leur sont géographiquement proches, plus l'autocorrélation spatiale est élevée.

Ce chapitre définit l'autocorrélation spatiale, puis décrit les indices d'autocorrélation spatiale au niveau global et local : principes, propriétés, mise en œuvre pratique avec R et interprétation de leur significativité.

R La lecture préalable des chapitres 1 : "Analyse spatiale descriptive" et 2 : "Codifier la structure de voisinage" est recommandée.

Très souvent, les variables pour lesquelles on dispose d'informations géolocalisées se caractérisent par des dépendances spatiales qui sont d'autant plus fortes que les localisations sont plus proches. Ainsi, l'accès de plus en plus fréquent à des données spatialisées permet de mieux prendre en compte les interactions et les externalités spatiales dans l'analyse des décisions économiques des agents. Une analyse des structures spatiales comprises dans les données est indispensable pour traiter, si cela s'avère nécessaire, la violation de l'hypothèse d'indépendance spatiale des variables. D'autre part, en termes d'interprétation, l'analyse de l'autocorrélation spatiale permet une analyse quantifiée de la structure spatiale du phénomène considéré. Les indices d'autocorrélation spatiale permettent de mesurer la dépendance spatiale entre les valeurs d'une même variable en différents endroits de l'espace.

3.1 Qu'est-ce que l'autocorrélation spatiale ?

L'autocorrélation mesure la corrélation d'une variable avec elle-même, lorsque les observations sont considérées avec un décalage dans le temps (autocorrélation temporelle) ou dans l'espace (autocorrélation spatiale). On définit l'autocorrélation spatiale comme la corrélation, positive ou négative, d'une variable avec elle-même du fait de la localisation spatiale des observations. Cette autocorrélation spatiale peut, d'une part, être le résultat de processus inobservés ou difficilement quantifiables qui associent des localisations différentes et qui, de ce fait, se traduisent par une structuration spatiale des activités : des phénomènes d'interaction (entre les décisions des agents par exemple) ou de diffusion (comme les phénomènes de diffusion technologique) dans l'espace sont autant de phénomènes qui peuvent produire de l'autocorrélaton spatiale. D'autre part, dans le contexte de la spécification de modèles économétriques, la mesure de l'autocorrélation spatiale peut être envisagée comme un outil de diagnostic et de détection d'une mauvaise spécification (variables omises spatialement corrélées, erreurs sur le choix de l'échelle à laquelle le phénomène spatial est analysé, etc.)

D'un point de vue statistique, de nombreuses analyses (analyse des corrélations, régressions linéaires, etc.) reposent sur l'hypothèse d'indépendance des variables. Lorsqu'une variable est spatialement autocorrélée, l'hypothèse d'indépendance n'est plus respectée, remettant ainsi en cause la validité des hypothèses sur la base desquelles ces analyses sont menées. D'autre part, l'analyse de l'autocorrélation spatiale permet une analyse quantifiée de la structure spatiale du phénomène étudié.
On insistera sur le fait que structure spatiale et autocorrélation spatiale ne peuvent pas exister indépendamment l'une de l'autre (TIEFELSDORF 1998) :

— on désigne par structure spatiale l'ensemble des liens grâce auxquels le phénomène autocorrélé va se diffuser ;
— sans la présence d'un processus autocorrélé significatif, la structure spatiale ne peut être empiriquement observée.

La distribution spatiale observée est alors considérée comme la manifestation du processus spatial sous-jacent.

3.1.1 Observation empirique de l'autocorrélation spatiale

En présence d'autocorrélation spatiale, on observe que la valeur d'une variable pour une observation est liée aux valeurs de cette même variable pour les observations voisines.

— L'autocorrélation spatiale est positive lorsque des valeurs similaires de la variable à étudier se regroupent géographiquement.

— L'autocorrélation spatiale est négative lorsque des valeurs dissemblables de la variable à étudier se regroupent géographiquement : des lieux proches sont plus différents que des lieux éloignés. On retrouve généralement ce type de situation en présence de concurrence spatiale.
— En l'absence d'autocorrélation spatiale, on peut considérer que la répartition spatiale des observations est aléatoire.

Les indices d'autocorrélation spatiale permettent d'évaluer la dépendance spatiale entre les valeurs d'une même variable en différents endroits de l'espace et de tester la significativité de la structure spatiale identifiée. Pour la mettre en évidence, les indices prennent en compte deux critères :

— la proximité spatiale ;
— la ressemblance ou la dissemblance des valeurs de la variable pour les unités spatiales considérées.

Attention : si les données sont agrégées suivant un découpage qui ne respecte pas le phénomène sous-jacent, on surestimera ou sous-estimera la force du lien spatial.

On distingue la mesure de l'autocorrélation spatiale globale d'une variable dans un espace donné et celle de l'autocorrélation locale dans chaque unité de cet espace. Celle-ci mesure l'intensité et la significativité de la dépendance locale entre la valeur d'une variable dans une unité spatiale et les valeurs de cette même variable dans les unités spatiales voisines (plus ou moins proches, selon le critère de voisinage retenu).

3.1.2 Le diagramme de Moran

Le diagramme de Moran permet une lecture rapide de la structure spatiale. Il s'agit d'un nuage de points avec les valeurs de la variable y centrée en abscisse et les valeurs moyennes de la variable pour les observations voisines Wy en ordonnée (où W est la matrice de poids normalisée). Les deux propriétés y centrée et W normalisée impliquent que la moyenne empirique de Wy est égale à celle de y et donc à 0. On trace également la droite de régression linéaire de Wy en fonction de y et les droites d'équation $y = 0$ et $Wy = 0$ qui délimitent des quadrants.

Si les observations sont réparties aléatoirement dans l'espace, il n'y a pas de relation particulière entre y et Wy. La pente de la droite de régression linéaire est nulle, et les observations sont réparties uniformément dans chacun des quadrants. Si au contraire les observations présentent une structure spatiale particulière, la pente de la régression linéaire est non nulle puisqu'il existe une corrélation entre y et Wy. Chacun des quadrants définis par $y = 0$ et $Wy = 0$ correspond à un type d'association spatiale particulier (figures 3.1 et 3.2).

— Les observations situées en haut à droite (quadrant 1) présentent des valeurs de la variable plus élevées que la moyenne, dans un voisinage qui leur ressemble (autocorrélation spatiale positive et valeur de l'indice élevé ; structure high-high).
— En bas à gauche (quadrant 3), les observations présentent des valeurs de la variable plus faibles que la moyenne, dans un voisinage qui leur ressemble (autocorrélation spatiale positive et valeur de l'indice faible ; structure low-low).
— Les observations situées en bas à droite (quadrant 2) ont des valeurs de la variable plus élevées que la moyenne dans un voisinage qui ne leur ressemble pas (autocorrélation spatiale négative et valeur de l'indice élevé ; structure high-low).
— En haut à gauche (quadrant 4), les observations présentent des valeurs de la variable plus basses que la moyenne dans un voisinage qui ne leur ressemble pas (autocorrélation spatiale négative et valeur de l'indice faible ; structure low-high).

La densité de points dans chacun des quadrants permet de visualiser la structure spatiale dominante. Le diagramme de Moran permet aussi de visualiser les points atypiques qui s'éloignent de cette structure spatiale.

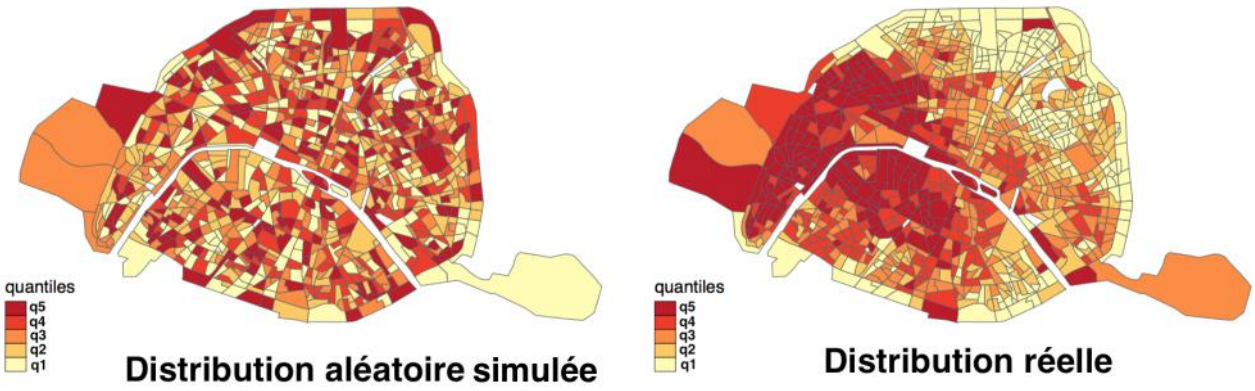


FIGURE 3.1 – Illustration, sur les Iris parisiens, de l'écart entre une distribution aléatoire et une distribution autocorrélée spatialement
Source : *Insee, Revenus Fiscaux Localisés 2010*

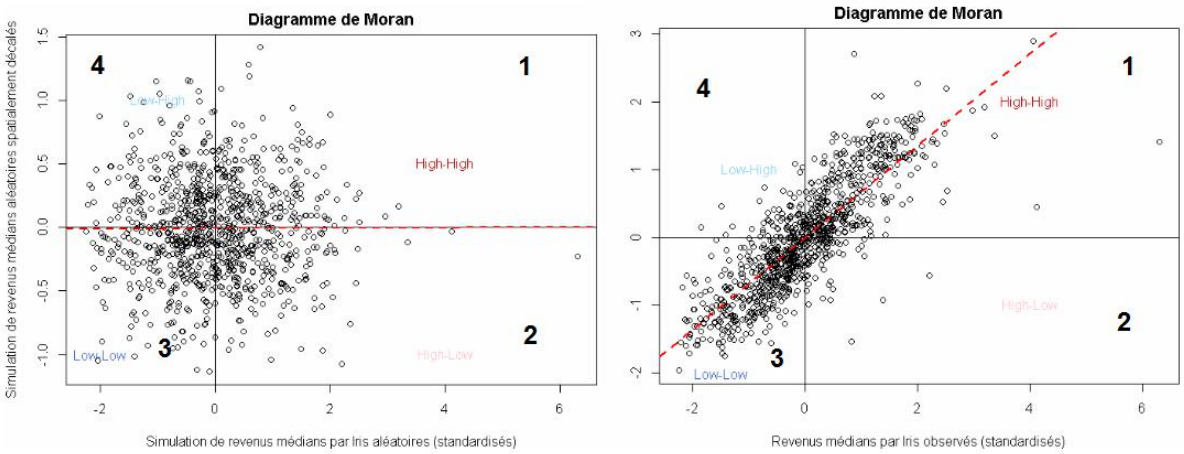


FIGURE 3.2 – Diagramme de Moran des revenus médians par Iris standardisés et d'une simulation de répartition aléatoire des revenus médians par Iris, pour les Iris parisiens
Source : *Insee, Revenus Fiscaux Localisés 2010*

Pour comprendre la façon dont l'autocorrélation spatiale est visible sur le diagramme de Moran, on a simulé une autocorrélation spatiale croissante des revenus par Iris (figures 3.3 et 3.4). Le paramètre ρ qui définit l'autocorrélation spatiale correspond à la pente du diagramme de Moran. À part pour les valeurs extrêmes, il est difficile d'identifier le signe et la force de l'autocorrélation spatiale en regardant simplement les cartes des différentes valeurs. En revanche, les diagrammes de Moran permettent d'identifier clairement les différents cas de figure.

3.2 Mesurer la dépendance spatiale globale

3.2.1 Indices d'autocorrélation spatiale

Lorsque le diagramme de Moran met en avant une structure spatiale particulière, le calcul des indices d'autocorrélation spatiale a pour objectif de répondre à deux questions :

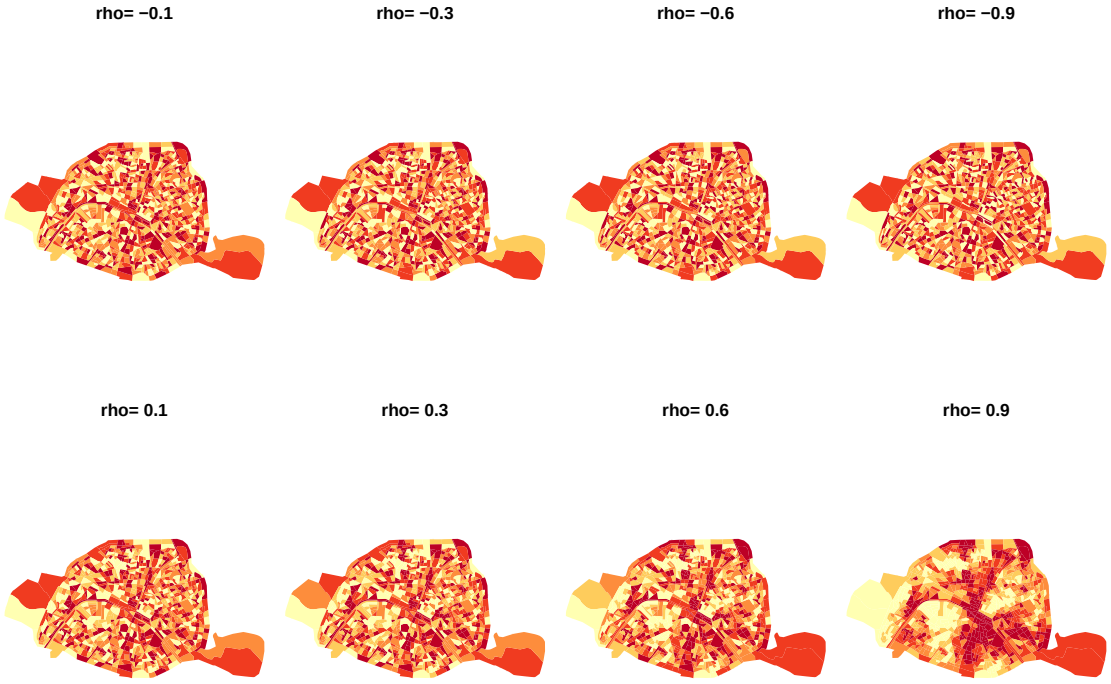


FIGURE 3.3 – Simulation d'une autocorrélation spatiale croissante des revenus par Iris parisiens
Source : *Insee, Revenus Fiscaux Localisés 2010*

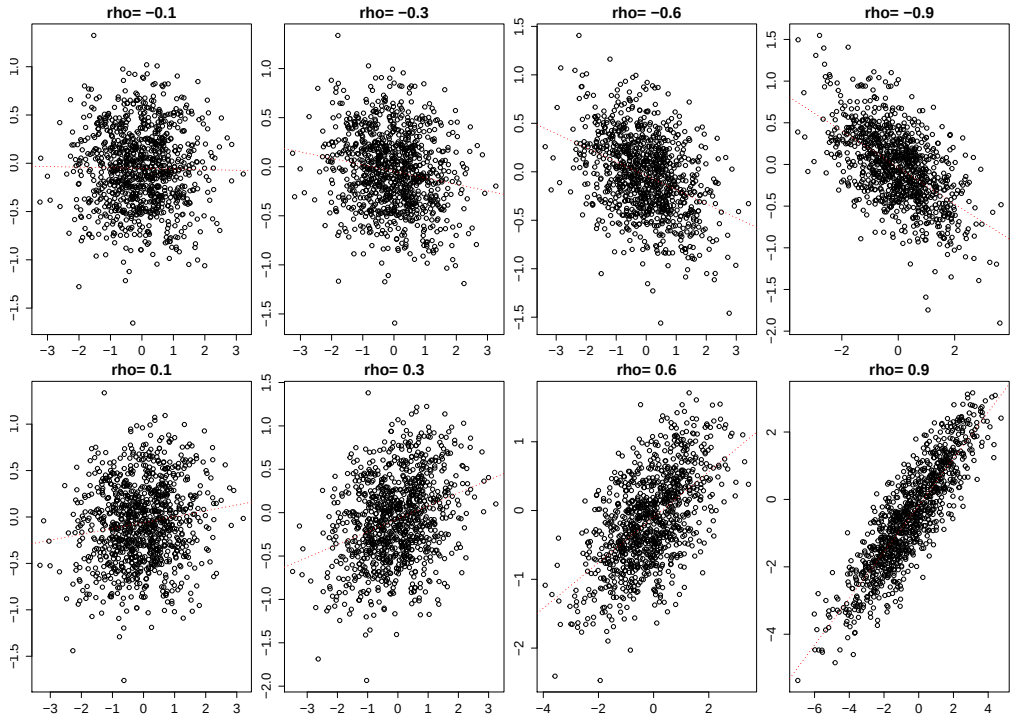


FIGURE 3.4 – Diagrammes de Moran associés aux simulations de revenus autocorrélés, pour les Iris parisiens
Source : *Insee, Revenus Fiscaux Localisés 2010*

— Les valeurs prises par les observations voisines auraient-elles pu être aussi comparables (ou aussi dissemblables) par le simple fait du hasard ?
— Si tel n'est pas le cas, il y a de l'autocorrélation spatiale : quels en sont le signe et la force ?

Répondre à la première question revient à tester l'hypothèse d'absence d'autocorrélation spatiale pour une variable brute y.
— H_0 : absence d'autocorrélation spatiale
— H_1 : présence d'autocorrélation spatiale
Pour mener à bien ce test, il faut préciser quelle est la distribution de la variable d'intérêt y, en l'absence d'autocorrélation spatiale (sous H_0). Dans ce contexte, l'inférence statistique est généralement menée en considérant l'une ou l'autre des deux hypothèses suivantes :

Hypothèse de normalité : chacune des valeurs de la variable, soit y_i, est le résultat d'un tirage indépendant dans la **distribution normale propre à chaque zone géographique i sur laquelle est mesurée cette variable**.

Hypothèse de randomisation : l'inférence sur le I de Moran est généralement menée sous l'hypothèse de randomisation. L'estimation de la statistique obtenue à partir des données est comparée avec **la distribution de celle obtenue en réordonnant au hasard (permutations) les données**. L'idée est simplement que si l'hypothèse nulle est vraie, alors toutes les combinaisons possibles des données sont équiprobables. Les données observées sont alors seulement l'une des réalisations parmi toutes celles également possibles. Dans le cas de l'autocorrélation spatiale, l'hypothèse nulle est toujours qu'il n'y a pas d'association spatiale et l'on affecte au hasard les valeurs de la variable aux unités spatiales afin de calculer la statistique du test. Si l'hypothèse nulle est rejetée, c'est-à-dire s'il y a de l'autocorrélation spatiale, on peut alors calculer l'intervalle de valeurs qui encadre l'indice d'autocorrélation spatiale et répondre ainsi à la question sur le signe et la force de l'autocorrélation spatiale : plus cet indice se rapproche de 1 en valeur absolue et plus la corrélation est élevée (cet intervalle dépend de la matrice de poids et peut parfois varier en dehors de l'intervalle $[-1;1]$, d'où l'intérêt de calculer les bornes de l'intervalle).

De manière très générale, les indices d'autocorrélation spatiale permettent de caractériser la corrélation entre les mesures géographiquement voisines d'un phénomène mesuré. Si l'on désigne par WY le vecteur des moyennes de la variable Y (où W désigne la matrice de pondération) dans le voisinage de chaque unité spatiale, les indices d'autocorrélation spatiale se mettent sous la forme :

$$Corr(Y,WY) = \frac{Cov(Y,WY)}{\sqrt{Var(Y).Var(WY)}} \tag{3.1}$$

À partir de cette formulation très générale, pour des variables quantitatives, deux indices sont principalement utilisés pour tester la présence d'autocorrélation spatiale : l'indice de Moran et l'indice de Geary. Le premier considère les variances et covariances en prenant en compte la différence entre chaque observation et la moyenne de toutes les observations. L'indice de Geary, lui, prend en compte la différence entre les observations voisines. Dans la littérature, l'indice de Moran est souvent préféré à celui de Geary en raison d'une stabilité générale plus grande (voir notamment UPTON et al. 1985).

Indice de Moran

$$I_W = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2} \quad i \neq j \tag{3.2}$$

— H_0 : Les voisins ne **co-varient** pas d'une façon particulière.
— $I_W > 0$ => autocorrélation spatiale positive.

Indice de Geary

$$c_W = \frac{n-1}{2\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - y_j)^2}{\sum_i (y_i - \bar{y})^2} \quad i \neq j \tag{3.3}$$

— H_0 : Les **différences** entre voisins n'ont pas de structure particulière.

— $c_W < 1$ => autocorrélation spatiale positive.

Selon la distribution retenue pour la variable en l'absence d'autocorrélation spatiale, le calcul de la variance des indices est modifié. En revanche, les équations qui donnent les expressions de l'espérance des indices (3.4) et la statistique du test (3.5) restent identiques. Ces relations permettent ainsi d'évaluer la significativité de l'autocorrélation spatiale.

$$\mathbb{E}(I_W) = \mathbb{E}(c_w) = -\frac{1}{n-1} \tag{3.4}$$

$$\frac{I_W - \mathbb{E}(I_W)}{\sqrt{Var(I_W)}} \sim \frac{c_W - \mathbb{E}(c_W)}{\sqrt{Var(c_W)}} \sim \mathcal{N}(0,1) \tag{3.5}$$

La mesure de l'autocorrélation spatiale reposant sur une comparaison de la valeur d'une variable pour un individu avec celle de ses voisins, la définition du voisinage va donc avoir un effet non négligeable sur la mesure de l'autocorrélation spatiale. Comme cela a été explicité dans le chapitre 2 "Codifier la structure de voisinage", plus le voisinage envisagé est large, plus on considère un nombre élevé de voisins, et plus la probabilité que leur moyenne se rapproche de la moyenne totale de la population va augmenter, ce qui risque de conduire à une valeur relativement faible de l'autocorrélation spatiale.

Un changement d'échelle peut également avoir des implications sur la mesure de l'autocorrélation spatiale. On désigne par MAUP (Modifiable Areal Unit Problem ; OPENSHAW et al. 1979b) 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. Selon OPENSHAW 1984, le MAUP est une combinaison de deux problèmes distincts mais proches :

— le problème de l'échelle 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 ;
— le problème de l'agrégation (ou de zonage) 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.

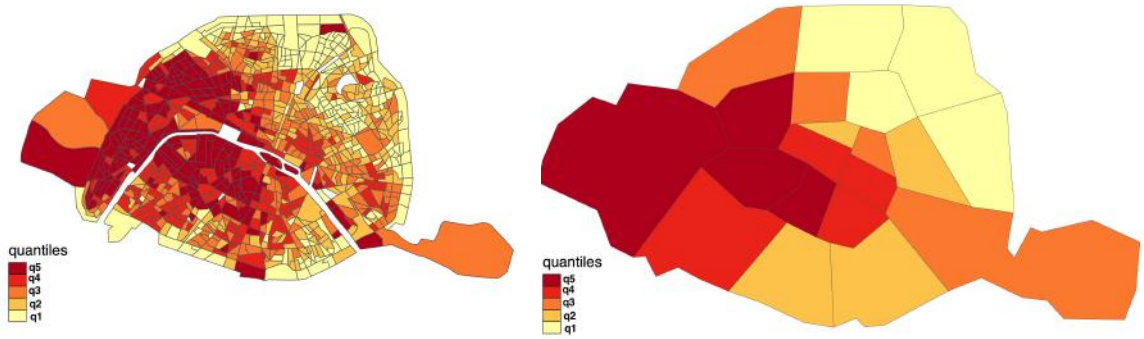
■ **Exemple 3.1 — Autocorrélation spatiale des revenus médians à Paris.** Quelle est l'intensité de l'autocorrélation spatiale des revenus parisiens ? Est-elle significative ? Dans quelle mesure dépend-elle de la spécification des relations spatiales (type de voisinage, échelle d'agrégation) ?

Source	I_W	c_w	p value	H0	bornes de I_W
Revenu : répartition observée	0.68	0.281	3.10^{-6}	rejetée	[-1.06,1.06]
Revenu : répartition aléatoire simulée	0.0027	1.0056	> 0.5	acceptée	[-1.06,1.06]

TABLE 3.1 – Indices de Moran et Geary des revenus médians des Iris parisiens : distribution réelle et simulée
Source : *Insee, Revenus Fiscaux Localisés 2010*

Type de voisinage	I_W	p value	H0
QUEEN	0.68	3.10^{-6}	rejetée
ROOK	0.57	2.10^{-6}	rejetée
1NN	0.30	0.07	rejetée
3NN	0.58	9.10^{-6}	rejetée
Delaunay	0.57	6.10^{-7}	rejetée

TABLE 3.2 – Indices de Moran et Geary des revenus médians des Iris parisiens en fonction de la définition du voisinage (voir chapitre 2 : "Codifier la structure de voisinage")
Source : *Insee, Revenus Fiscaux Localisés 2010*



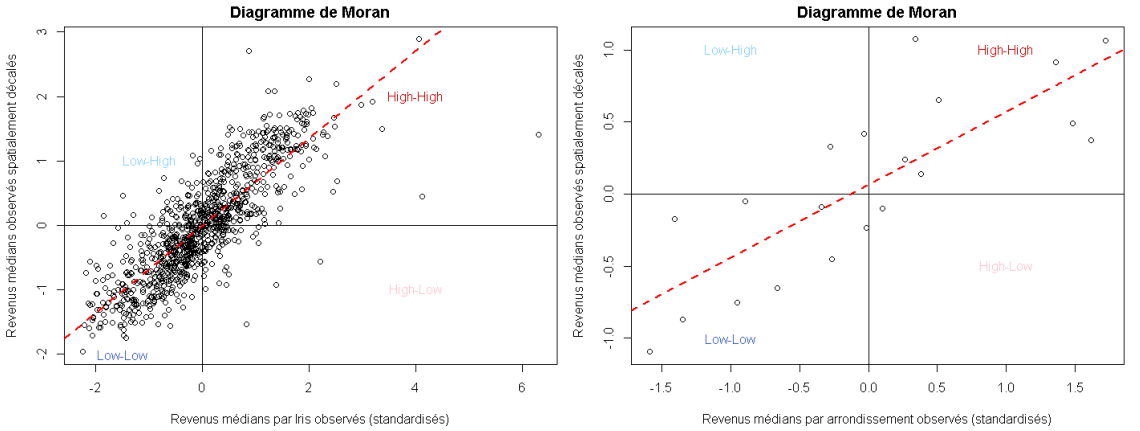
(a) à l'Iris

(b) à l'arrondissement

FIGURE 3.5 – Agrégation des revenus à Paris
Source : *Insee, Revenus Fiscaux Localisés 2010*

échelle d'agrégation	I_W	p value	H0	bornes de I_W
Iris	0.68	3.10^{-6}	rejetée	[-1.06,1.06]
Arrondissement	0.51	$< 9.10^{-9}$	rejetée	[-0.53,1.01]

TABLE 3.3 – Valeur et significativité du I de Moran en fonction de l'échelle d'agrégation choisie
Source : *Insee, Revenus Fiscaux Localisés 2010*



(a) à l'IRIS (b) à l'arrondissement

FIGURE 3.6 – Diagrammes de Moran pour la distribution des revenus à Paris
Source : *Insee, Revenus Fiscaux Localisés 2010*

Dans cet exemple (voir tables 3.1, 3.2, 3.3 et figures 3.5, 3.6), quelle que soit la définition du voisinage ou l'échelle d'agrégation, l'autocorrélation spatiale des revenus parisiens est positive et significative. La force de l'autocorrélation spatiale varie légèrement selon le type de voisinage retenu (en particulier, dans cet exemple, considérer uniquement les plus proches voisins diminue un peu la force de l'autocorrélation spatiale mesurée). ■

Application avec R

Le package *spdep* permet de calculer les indices d'autocorrélation spatiale et leur significativité grâce aux fonctions `moran.test` et `geary.test`.
Par défaut, la distribution de la variable d'intérêt sous l'hypothèse nulle est obtenue par randomisation. L'argument `randomisation = FALSE` permet de supposer qu'il s'agit d'une distribution normale.

Encadré 3.2.1 — Si certaines entités n'ont pas de voisins. Pour que les fonctions du package *spdep* acceptent des matrices de poids dans lesquelles certaines entités n'ont pas de voisins, il est nécessaire de spécifier l'option : `zero.policy=TRUE`. Par défaut, la taille de la matrice est diminuée pour exclure les observations n'ayant pas de voisins. On peut spécifier le contraire avec l'option : `adjust.n=FALSE`. Dans ce cas, la valeur absolue de la statistique de test augmente, et la valeur absolue de son espérance et de sa variance diminuent (BIVAND et al. 2013a). De façon générale, les indices d'autocorrélation spatiale ont été développés en supposant que toutes les entités avaient des voisins, et les avis ne sont pas unanimes sur l'attitude à adopter lorsque tel n'est pas le cas.

Comme vu précédemment, il existe deux approches pour estimer la significativité de ces indices : une solution analytique qui s'appuie sur l'hypothèse de normalité et une solution de Monte Carlo qui s'appuie sur l'hypothèse de randomisation. La solution analytique (utilisée par la fonction `moran.test`) fait l'hypothèse que la statistique du test suit asymptotiquement une loi normale de moyenne 0 et de variance 1. Cela peut ne pas toujours s'avérer être la mesure la plus précise de la significativité car la convergence vers cette loi peut dépendre de l'arrangement des polygones. On peut utiliser à la place la fonction `moran.mc` qui permet de choisir le nombre de permutations pour calculer la distribution simulée du I de Moran. Comparer les seuils de significativité calculés à partir des fonctions `moran.mc` et `moran.test` permet de s'assurer de la robustesse des conclusions.

```
library(spdep)

#####################################
# Préparation des données  ##########
#####################################

#Extraction de la liste des voisins (au sens Queen par défaut)
iris75.nb <- poly2nb(iris75)
#Création de la matrice de poids
iris75.lw <- nb2listw(iris75.nb,zero.policy=TRUE)
#Calcul des revenus médians standardisés
iris75.data <- as.data.frame(iris75)
iris75.data$med_revenu_std <- scale(iris75.data$med_revenu)

######################
# Diagramme de MORAN
######################

moran.plot(iris75.data$med_revenu_std,iris75.lw,labels=FALSE,
 xlab='revenus medians par IRIS',ylab='moyenne des revenus médians par IRIS
      des voisins')

######################
# Test du I de Moran
######################

moran.test(iris75.data$med_revenu_std,iris75.lw,zero.policy=TRUE,
    randomisation=FALSE)

#Calcul des intervales du I de Moran :
moran.range <- function(lw) {
    wmat <- listw2mat(lw)
    return(range(eigen((wmat+t(wmat))/2)$values))
}

moran.range(iris75.lw)
```

3.2.2 Autocorrélation spatiale des variables catégorielles

Lorsque la variable d'intérêt n'est pas continue, mais catégorielle, on mesure le degré d'association locale grâce à une analyse des statistiques des *join count* (ZHUKOV 2010).

Pour illustrer le calcul de ces statistiques, on considère une variable binaire qui représente deux couleurs, Blanc (B) et Noir (N) de sorte qu'une liaison puisse être qualifiée de Blanc-Blanc, Noir-Noir ou Blanc-Noir. On observe :

- une autocorrélation spatiale positive si le nombre de liaisons Blanc-Noir est significativement **inférieur** à ce que l'on aurait obtenu à partir d'une répartition spatiale aléatoire ;
- une autocorrélation spatiale négative si le nombre de liaisons Blanc-Noir est significativement **supérieur** à ce que l'on aurait obtenu à partir d'une répartition spatiale aléatoire ;

— aucune autocorrélation spatiale si le nombre de liaisons Blanc-Noir est approximativement **identique** à ce que l'on aurait obtenu à partir d'une répartition spatiale aléatoire.

S'il y a n observations, n_b observations blanches et $n_n = n - n_b$ observations noires, la probabilité d'obtenir une observation blanche est : $P_b = \frac{n_b}{n}$ et la probabilité d'obtenir une observation noire est : $P_n = 1 - P_b$.
En l'absence d'autocorrélation spatiale, les probabilités d'obtenir des observations d'une même couleur dans deux cellules voisines sont : $P_{bb} = P_b * P_b = P_b^2$ et $P_{nn} = P_n * P_n = (1 - P_b)^2$.
La probabilité d'obtenir des observations de couleur différente dans deux cellules voisines est : $P_{bn} = P_b * (1 - P_b) + (1 - P_b) * P_b = 2P_b * (1 - P_b)$.

Comme $\frac{1}{2}\sum_i \sum_j w_{ij}$ mesure le nombre de liaisons existantes, sous l'hypothèse d'une répartition spatiale aléatoire des observations, on peut écrire :

$$\begin{aligned}
\mathbb{E}[bb] &= \frac{1}{2}\sum_i \sum_j w_{ij} P_b^2 \\
\mathbb{E}[nn] &= \frac{1}{2}\sum_i \sum_j w_{ij} (1 - P_b)^2 \\
\mathbb{E}[bn] &= \frac{1}{2}\sum_i \sum_j w_{ij} 2P_b * (1 - P_b)
\end{aligned} \tag{3.6}$$

Si l'on désigne par $y_i = 1$ lorsque l'observation est de couleur noire et par $y_i = 0$ dans le cas contraire (couleur blanche), les contre-parties empiriques (valeurs observées) de ces espérances mathématiques peuvent s'écrire :

$$\begin{aligned}
nn &= \frac{1}{2}\sum_i \sum_j w_{ij} y_i y_j \\
bb &= \frac{1}{2}\sum_i \sum_j w_{ij} (1 - y_i)(1 - y_j) \\
bn &= \frac{1}{2}\sum_i \sum_j w_{ij} (y_i - y_j)^2
\end{aligned} \tag{3.7}$$

Dans ce cas, la statistique de test permettant d'évaluer la significativité de l'autocorrélation spatiale repose sur l'hypothèse qu'en l'absence d'autocorrélation spatiale, les statistiques de *join count* (*bb, nn* et *bn*) suivent une loi normale. On peut alors écrire :

$$\frac{bn - \mathbb{E}(bn)}{\sqrt{Var(bn)}} \sim \mathcal{N}(0,1) \quad \frac{bb - \mathbb{E}(bb)}{\sqrt{Var(bb)}} \sim \mathcal{N}(0,1) \quad \frac{nn - \mathbb{E}(nn)}{\sqrt{Var(nn)}} \sim \mathcal{N}(0,1) \tag{3.8}$$

■ **Exemple 3.2 — Statistiques *join count* pour l'emploi des individus parisiens.** [1]

On considère la variable binaire qui vaut 1 si l'individu i est chômeur et 0 sinon. On cherche à déterminer si les chômeurs parisiens sont plus regroupés dans l'espace que s'ils étaient répartis aléatoirement. Les statistiques de *join count* permettent de répondre à cette question. À partir de la table 3.4 on peut constater que la localisation des chômeurs est significativement corrélée, tout comme l'est celle des actifs.

1. L'objectif de cet exemple n'est pas de détailler les résultats d'une étude économique, mais d'illustrer les techniques mises en œuvre. Il n'y a aucune interprétation à en tirer.

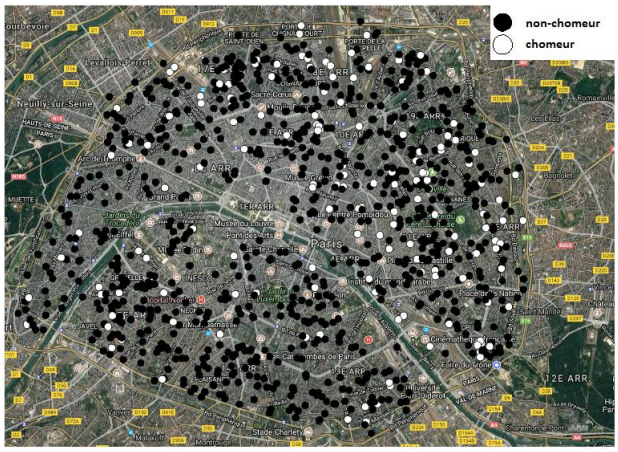


FIGURE 3.7 – Emploi d'un échantillon de 1000 individus parisiens
Source : *Insee, Revenus Fiscaux Localisés 2010*

Variable	p-value de la statistique joincount d'association spatiale	H0
Chômeur	$5.439.10^{-}3$	rejetée
Actif	$9.085.10^{-}5$	rejetée

TABLE 3.4 – Significativité de la statistique du joincount des chômeurs parisiens
Source : *Insee, Revenus Fiscaux Localisés 2010*

Application avec R

Cette statistique *joint count* est obtenue en mettant en œuvre la fonction `joincount.test` du package R *spdep*.

```
library(spdep)

#Conversion en facteur
menir10_d75_subset$chomage <- ifelse(menir10_d75_subset$ZCHOM>0,3,1)
chomage <- as.factor(menir10_d75_subset$chomage,levels=c("actif","chomeur")
    )

#Création des listes de voisins et matrices de poids
coordinates(menir10_d75_subset) <- c("PLG_X", "PLG_Y")
proj4string(menir10_d75_subset)<- CRS( "+init=epsg:27572 +proj=lcc +lat_
    1=46.8 +lat_0=46.8 +lon_0=0 +k_0=0.99987742 +x_0=600000 +y_0=2200000 +a
    =6378249.2 +b=6356515 +towgs84=-168,-60,320,0,0,0,0 +pm=paris +units=m
    +no_defs")
menir10_d75_subset <- spTransform (menir10_d75_subset, CRS ("+init=epsg
    :2154") )

menir75.nb<-  knn2nb(knearneigh(menir10_d75_subset,k=2))

#Mise en oeuvre du test
joincount.test(chomage,listw2U(nb2listw(menir75.nb)))
```

Dans le cas de plusieurs catégories, la fonction `joincount.multi` du package *spdep* permet

de tester la significativité, selon le même principe, de l'association spatiale de différentes variables.

■

3.3 Mesurer la dépendance spatiale locale

Les statistiques globales font l'**hypothèse de stationnarité du processus spatial** : l'autocorrélation spatiale serait la même dans tout l'espace. Or cette hypothèse est d'autant moins réaliste que le nombre d'observations est élevé.

3.3.1 Indice de Getis et Ord

Getis et Ord (GETIS et al. 1992) proposent un indicateur permettant de détecter les dépendances spatiales locales qui n'apparaissent pas dans l'analyse globale.

Indicateur de Getis et Ord

$$G_i = \frac{\sum_j w_{ij} y_j}{\sum_j w_{ij}} \tag{3.9}$$

$G_i > 0$ indique un regroupement de valeurs plus élevées que la moyenne.
$G_i < 0$ indique un regroupement de valeurs plus basses que la moyenne.

On peut tester la significativité de l'indicateur de Getis et Ord en faisant l'hypothèse, en l'absence de dépendance spatiale locale, d'une distribution normale.

$$z(G_i) = \frac{G_i - \mathbb{E}(G_i)}{\sqrt{Var(G_i)}} \sim \mathcal{N}(0,1) \tag{3.10}$$

Application avec R

La fonction `localG` du package *spdep* permet d'utiliser cet indicateur.

3.3.2 Indicateurs d'autocorrélation spatiale locale

Anselin (ANSELIN 1995) développe les notions introduites par Getis et Ord en définissant des *indicateurs d'autocorrélation spatiale locale*. Ceux-ci doivent permettre de mesurer l'intensité et la significativité de la dépendance locale entre la valeur d'une variable dans une unité spatiale et les valeurs de cette même variable dans les unités spatiales environnantes. Plus précisément, ces indicateurs permettent de :

— détecter les regroupements significatifs de valeurs identiques autour d'une localisation particulière (clusters) ;
— repérer les zones de non-stationnarité spatiale, qui ne suivent pas le processus global.

Les indicateurs de Getis et Ord ne répondent qu'au premier de ces deux objectifs. Pour être considérés comme des mesures locales d'association spatiale (LISA ; *Local Indicators of Spatial Association*) telles qu'elles ont été définies par Anselin, ces indicateurs doivent vérifier les deux propriétés suivantes :

— pour chaque observation, ils indiquent l'intensité du regroupement de valeurs similaires (ou de tendance opposée) autour de cette observation ;
— la somme des indices locaux sur l'ensemble des observations est proportionnelle à l'indice global correspondant.

Le LISA le plus couramment utilisé est le I de Moran local.

I de Moran local

$$I_i = (y_i - \bar{y}) \sum_j w_{ij}(y_j - \bar{y}) \tag{3.11}$$

$$I_W = constante * \sum_i I_i \tag{3.12}$$

$I_i > 0$ indique un regroupement de valeurs similaires (plus élevées ou plus faibles que la moyenne). $I_i < 0$ indique un regroupement de valeurs dissimilaires (par exemple des valeurs élevées entourées de valeurs faibles).

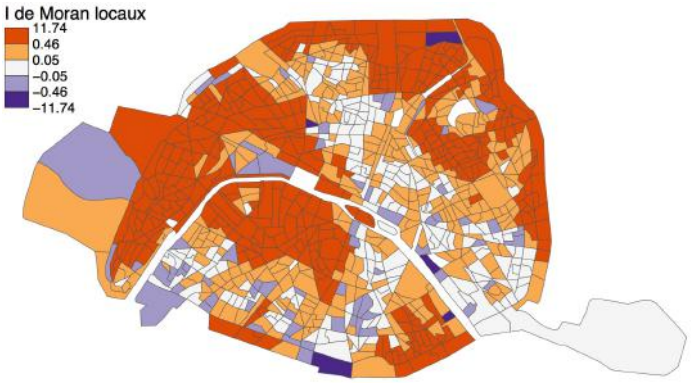


FIGURE 3.8 – Valeurs des I de Moran locaux, sur les Iris parisiens
Source : *Insee, Revenus Fiscaux Localisés 2010*

3.3.3 Significativité du I de Moran local

Des LISA significatifs correspondent à des regroupements de valeurs similaires ou dissimilaires plus marqués que ce que l'on aurait pu observer à partir d'une répartition spatiale aléatoire. Ces regroupements peuvent correspondre aux quatre types de regroupements spatiaux décrits en 3.1 et identifiables sur le diagramme de Moran (high-high, low-low, high-low ou low-low). Le test de significativité de chaque indicateur d'association locale repose sur une statistique supposée suivre asymptotiquement une loi normale sous l'hypothèse nulle. En effet, si l'on admet que l'hypothèse de normalité est vérifiée, $z(I_i) = \frac{I - \mathbb{E}(I_i)}{\sqrt{Var(I_i)}} \sim \mathcal{N}(0,1)$.

Pour tester la validité de l'hypothèse de normalité des LISA sous l'hypothèse nulle, on simule plusieurs répartitions aléatoires dans l'espace de la variable d'intérêt puis on calcule les indicateurs locaux associés à ces simulations.

En reprenant l'exemple des revenus parisiens, on observe (figure 3.10) que les quantiles extrêmes de la distribution des I locaux sont plus élevés que ceux d'une distribution normale. Les *p-values* calculées sous l'hypothèse de normalité devront donc être utilisées avec précaution. En effet, Anselin (ANSELIN 1995) montre, à partir de simulations (figure 3.11), **qu'en présence d'autocorrélation spatiale globale, l'hypothèse de normalité des I_i n'est plus vérifiée.**

D'autre part, le test de significativité des LISA soulève un problème que l'on rencontre à chaque fois que l'on effectue des comparaisons multiples. En effet, lorsque plusieurs tests statistiques sont réalisés simultanément à partir du même jeu de données, le risque global d'erreur de première espèce (probabilité de rejeter à tort l'hypothèse nulle) s'accroît. La répétition à chaque test du risque d'obtenir un résultat significatif par hasard augmente le risque global de conclure

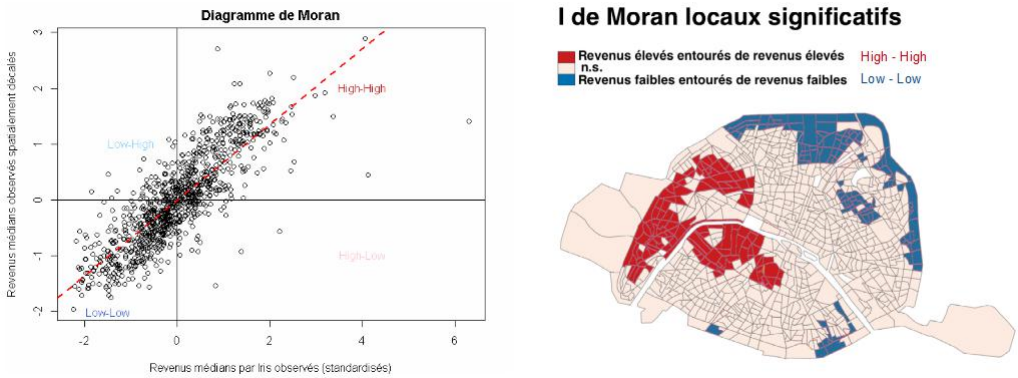


FIGURE 3.9 – I de Moran locaux significatifs, sur les Iris parisiens
Source : *Insee, Revenus Fiscaux Localisés 2010*

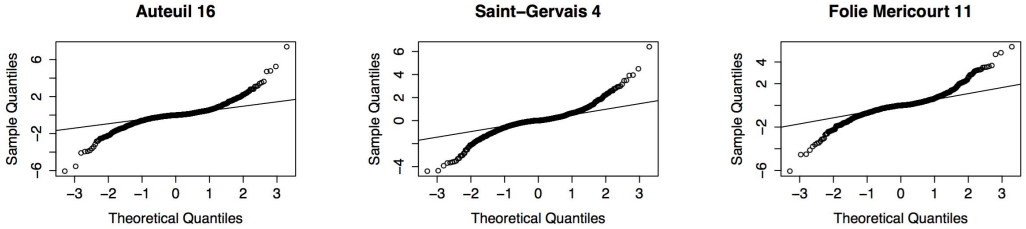


FIGURE 3.10 – Test de l'hypothèse de normalité de la distribution des I de Moran locaux sur trois Iris parisiens
Source : *Insee, Revenus Fiscaux Localisés 2010*

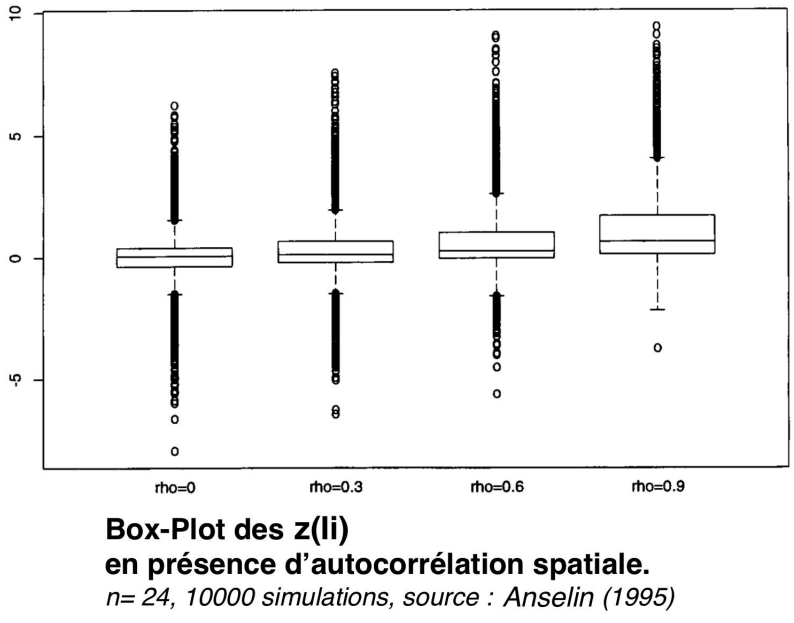


FIGURE 3.11 – Distribution des I de Moran locaux en présence d'autocorrélation spatiale globale
Source : *Insee, Revenus Fiscaux Localisés 2010*

à tort à la significativité de l'indice local. Ainsi, dans notre cas, on conclura à **l'existence** d'une autocorrélation spatiale locale si **au moins un** indice d'autocorrélation spatiale locale est significatif parmi tous les indices de la zone d'étude. S'il y a 100 indices d'autocorrélation spatiale locaux, on multiplie par 100 le risque d'en détecter au moins un significatif à tort (formule précisée en encadré 3.3.1). Il y a inflation du risque α (erreur de type I) : on augmente le risque de conclure à tort à l'existence d'une autocorrélation spatiale locale (ANSELIN 1995 ; ORD et al. 1995).

Différentes méthodes ont été développées pour éviter cette inflation du risque α lorsque plusieurs comparaisons statistiques sont nécessaires. Nous en détaillons quelques-unes dans ce qui suit.
Soit α le seuil de significativité retenu pour chaque indice local.

Encadré 3.3.1 — Méthode de Bonferroni : la méthode historique. La probabilité de ne pas rejeter à tort H_0 est $1-\alpha$ par polygone, donc $(1-\alpha)^n$ pour toute la zone, avec n le nombre de polygones.
La probabilité de rejeter au moins une fois à tort H_0 est $\alpha^* = 1-(1-\alpha)^n \approx n\alpha$.
Si l'on veut maintenir le risque global approximativement à α, on peut donc retenir $\alpha' \approx \frac{\alpha^*}{n}$ comme seuil de chaque test individuel. Ainsi pour $\alpha = 0.05$, un regroupement est significatif si sa p-value vaut $\frac{0.05}{n}$.
Le logiciel R permet de mettre en place cette méthode avec l'option `method='bonferroni'` de la fonction `p.adjust`.
On considère que cette méthode ne donne de bons résultats que lorsque le nombre de tests réalisés est petit. Dans le cas des LISA elle est un peu trop restrictive et l'on s'expose au risque, vu le nombre de comparaisons, de ne pas détecter certains LISA significatifs.

Encadré 3.3.2 — Méthode d'ajustement de Holm : permet de détecter l'existence d'un cluster spatial. La méthode d'ajustement de (HOLM 1979) prend en compte le fait que si parmi les n polygones, k sont vraiment des clusters spatiaux significatifs, la probabilité de rejeter à tort H_0 sur toute la zone n'est pas $(1-\alpha)^n$ mais $(1-\alpha)^{n-k}$ où α est le seuil de significativité souhaité.

La méthode de Holm classe les p-values de α_1 la plus faible à α_n la plus élevée.
Si $\alpha'_1 \sim n\alpha_1 < \alpha$, *i.e.* $\alpha_1 < \frac{\alpha}{n}$, on considère que cet indice local est effectivement significatif puisqu'il remplit le critère le plus restrictif. On regarde alors si $\alpha_2 < \frac{\alpha}{n-1}$, et ainsi de suite jusqu'à tester si $\alpha_k < \frac{\alpha}{n-k+1}$.
Le logiciel R permet de mettre en place cette méthode avec l'option `method='holm'` de la fonction `p.adjust`.
La méthode d'ajustement de Holm conduit à un plus grand nombre de clusters significatifs que la méthode de Bonferroni. Elle lui est donc le plus souvent préférée. Cependant, cette méthode se concentre aussi sur la détection **de l'existence d'au moins un cluster dans toute la zone**.

Encadré 3.3.3 — Méthode du False Discovery Rate : permet de localiser les clusters spatiaux. La méthode du False Discovery Rate (FDR) a été introduite par BENJAMINI et al. 1995. Avec cette méthode, le risque de juger - à tort - un cluster comme significatif est plus élevé, mais inversement le risque de juger - à tort - un cluster comme non significatif est plus faible. CALDAS DE CASTRO et al. 2006 prouvent l'intérêt de cette méthode pour **localiser** les clusters spatiaux significatifs.
La méthode du FDR classe les p-values de α_1 la plus faible à α_n la plus élevée.

Soit k le plus grand entier tel que $\alpha_k \leq \frac{k}{n}\alpha$. Benjamini et Hochberg expliquent qu'on peut rejeter l'hypothèse nulle d'absence d'autocorrélation spatiale locale pour tous les clusters dont les p-values sont inférieures ou égales à α_k.
Le logiciel R permet de mettre en place cette méthode avec l'option `method='fdr'` de la fonction `p.adjust`.

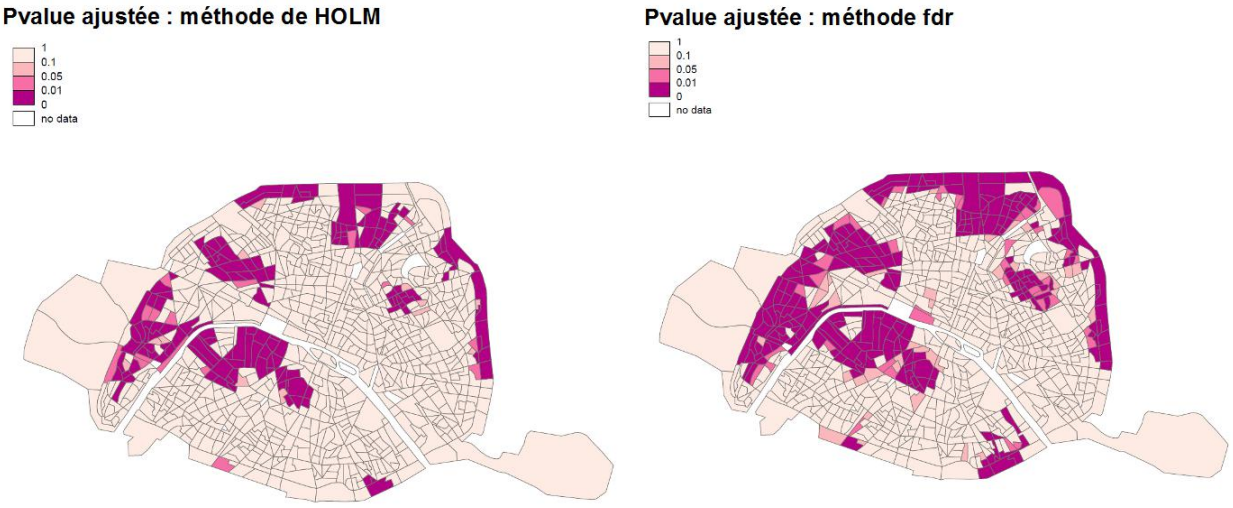


FIGURE 3.12 – Test de la significativité des I de Moran locaux, sur les Iris parisiens
Source : *Insee, Revenus Fiscaux Localisés 2010*

Dans l'exemple des revenus parisiens (figure 3.12), on observe bien que l'ajustement des p-values par la méthode de Holm conduit à moins de p-values significatives que l'ajustement par la méthode FDR. La méthode de Holm diminue en effet le risque de conclure à tort à l'**existence** d'une autocorrélation spatiale locale. En revanche, cette méthode augmente le risque de passer à côté d'un cluster local. Le choix de la méthode d'ajustement dépendra donc des objectifs de l'étude et des risques que l'on privilégie.

Application avec R

```
lisa_revenus<-  localmoran(iris75.data$med_revenu,iris75.lw,zero.policy=
    TRUE)

#Calcul des p-values ajustées
iris75.data.LISA$pvalue_ajuste<-
  p.adjust(iris75.data.LISA$pvalue_LISA,method='bonferroni')
```

3.3.4 Interprétation des indices locaux

En l'absence d'autocorrélation spatiale globale

Les LISA permettent de **détecter les zones où des valeurs similaires se regroupent de façon significative**. Il s'agit de zones où la structure spatiale locale est telle que les liens entre voisins sont particulièrement forts.

En présence d'autocorrélation spatiale globale

Les LISA **indiquent les zones qui influent particulièrement sur le processus global** (autocorrélation locale plus marquée que l'autocorrélation globale), **ou au contraire qui s'en dé-**

marquent (plus faible autocorrélation). Ainsi, dans l'exemple des revenus médians parisiens, on observe que la distribution des I de Moran locaux n'est pas centrée sur le I de Moran global (figure 3.13). Certaines zones ont une structure d'association spatiale significativement différente du processus global.

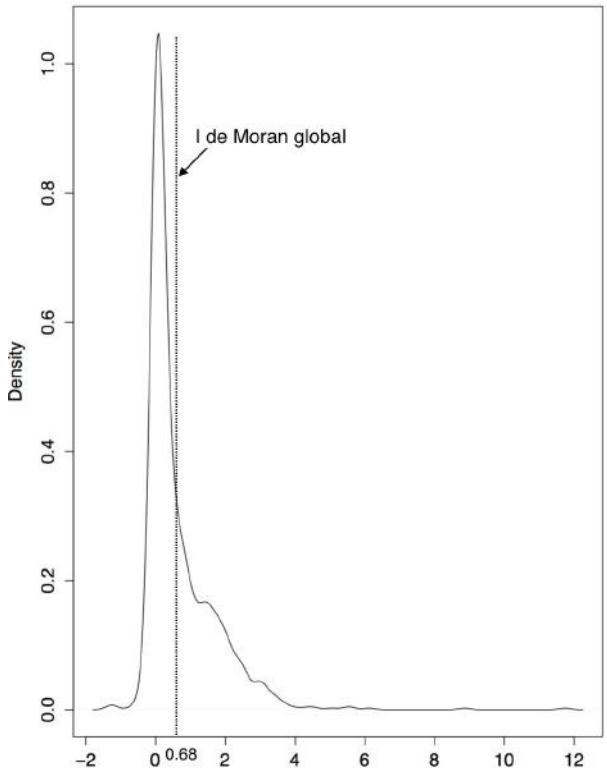


FIGURE 3.13 – Distribution des I de Moran locaux des revenus médians par Iris parisiens
Source : *Insee, Revenus Fiscaux Localisés 2010*

Même ajustées, les p-values risquent d'être trop faibles, puisque la distribution des I_i s'éloigne de la normale. Plus l'autocorrélation globale augmente, plus le nombre de valeurs extrêmes augmente. Les LISA élevés peuvent donc difficilement être interprétés comme des regroupements significatifs de valeur similaire. Dans ce cas, on interprète les LISA comme indicateurs d'une certaine **instabilité locale**.

3.4 Indices spatio-temporels

Il n'est pas rare qu'une base de données géolocalisée comporte des observations relevées à différents moments dans le temps, comme c'est le cas avec les bases de données recensant les transactions sur les biens immobiliers. Il peut être intéressant de comprendre comment un phénomène localisé s'est diffusé et a évolué dans l'espace et dans le temps et comment cela peut être lié aux conditions de l'environnement qui l'entoure. Dans ce cas, il faut être en mesure d'évaluer comment les structures spatiales sous-jacentes se modifient à différentes périodes de temps. Sur des données spatio-temporelles, l'exploration graphique préalable des données en coupe (I de Moran classique) peut permettre d'étudier l'existence et l'évolution des tendances de regroupement ou de dispersion qui sont statistiquement significativement différentes des modèles aléatoires.

De nombreux développements récents montrent un intérêt croissant pour l'analyse des données spatio-temporelles dans de nombreux domaines de recherche tels que la physique, la météorologie, l'économie et les études environnementales. En prolongeant l'indice de Moran pour y inclure des attributs temporels, il devient possible de calculer des indices globaux et localisés qui tiennent compte simultanément des autocorrélations spatiale et temporelle. Cela peut également se faire à partir de matrices de pondérations spatio-temporelles. Les travaux de MARTIN et al. 1975, WANG et al. p.d., LÓPEZ-HERNÁNDEZ et al. 2007 proposent des extensions du I de Moran traditionnellement utilisé pour mesurer la dépendance spatiale, pour calculer un I de Moran spatio-temporel. CHEN et al. 2013 développent une approche analytique améliorée fondée sur le I de Moran traditionnel et qui repose sur la stationnarité des données dans le temps. Comme le remarquent LEE et al. 2017, les séries temporelles géolocalisées sont généralement non stationnaires. Lorsque cette hypothèse n'est pas respectée, l'indice de Moran spatio-temporel proposé par CHEN et al. 2013 peut être fallacieux. LEE et al. 2017 proposent de contourner cette difficulté en appliquant une correction des fluctuations autour de la tendance (detrended fluctuation analysis, DFA) et suggèrent une nouvelle méthode de calcul de cet indice.

Conclusion

Les indices d'autocorrélation spatiale sont des outils de statistique exploratoire qui permettent de mettre en évidence l'existence d'un phénomène spatial significatif. Les sections 2 et 3 présentent des méthodes différentes pour prendre en compte ce phénomène spatial, au niveau global ou local, pour des variables quantitatives ou qualitatives. Il est important de savoir si l'autocorrélation est significative ou pas, mais également de mesurer la portée de l'autocorrélation afin de déterminer l'échelle de la dépendance spatiale. L'étude de l'autocorrélation spatiale est une étape indispensable avant d'envisager toute spécification des interactions spatiales dans un modèle approprié.

Références - Chapitre 3

ANSELIN, Luc (1995). « Local indicators of spatial association—LISA ». *Geographical analysis* 27.2, p. 93–115.

BENJAMINI, Yoav et Yosef HOCHBERG (1995). « Controlling the false discovery rate : a practical and powerful approach to multiple testing ». *Journal of the royal statistical society. Series B (Methodological)*, p. 289–300.

BIVAND, Roger S, Edzer PEBESMA et Virgilio GOMEZ-RUBIO (2013a). *Applied spatial data analysis with R*. T. 10. Springer Science & Business Media.

CALDAS DE CASTRO, Marcia et Burton H SINGER (2006). « Controlling the false discovery rate : a new application to account for multiple and dependent tests in local statistics of spatial association ». *Geographical Analysis* 38.2, p. 180–208.

CHEN, S.-K et al. (2013). « Analysis on Urban Traffic Status Based on Improved Spatiotemporal Moran's I ». *Acta Physica Sinica* 62.14.

GETIS, Arthur et J Keith ORD (1992). « The analysis of spatial association by use of distance statistics ». *Geographical analysis* 24.3, p. 189–206.

HOLM, Sture (1979). « A simple sequentially rejective multiple test procedure ». *Scandinavian journal of statistics*, p. 65–70.

LEE, Jay et Shengwen LI (2017). « Extending moran's index for measuring spatiotemporal clustering of geographic events ». *Geographical Analysis* 49.1, p. 36–57.

LÓPEZ-HERNÁNDEZ, Fernando A et Coro CHASCO-YRIGOYEN (2007). « Time-trend in spatial dependence : Specification strategy in the first-order spatial autoregressive model ». *Estudios de Economia Aplicada* 25.2.

MARTIN, Russell L et JE OEPPEN (1975). « The identification of regional forecasting models using space : time correlation functions ». *Transactions of the Institute of British Geographers*, p. 95–118.

OPENSHAW, Stan (1984). *The modifiable areal unit problem*. T. CATMOG 38. GeoBooks, Norwich, England.

OPENSHAW, Stan et Peter TAYLOR (1979b). « A million or so correlation coefficients ». *Statistical methods in the spatial sciences*, p. 127–144.

ORD, J Keith et Arthur GETIS (1995). « Local spatial autocorrelation statistics : distributional issues and an application ». *Geographical analysis* 27.4, p. 286–306.

TIEFELSDORF, Michael (1998). « Modelling spatial processes : The identification and analysis of spatial relationships in regression residuals by means of Moran's I (Germany) ». Thèse de doct. Université Wilfrid Laurier.

UPTON, Graham, Bernard FINGLETON et al. (1985). *Spatial data analysis by example. Volume 1 : Point pattern and quantitative data.* John W & Sons Ltd.

WANG, Y. F. et H. L. HE. « Spatial Data Analysis Method ». *Science Press, Beijing, China*.

ZHUKOV, Yuri M (2010). « Applied spatial statistics in R, Section 2 ». *Geostatistics.[Online] Available : http ://www. people. fas. harvard. edu/˜ zhukov/Spatial5. pdf*.

4. Les configurations de points

JEAN-MICHEL FLOCH
Insee

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

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

Résumé

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

Comprendre et maîtriser des méthodes statistiques qui traitent ces informations individuelles et spatialisées permet de travailler sur des données qui sont aujourd'hui de plus en plus accessibles et recherchées car elles fournissent des analyses très précises sur les comportements des acteurs économiques (ELLISON et al. 2010 ; BARLET et al. 2013). Dans ce cadre d'analyse, plusieurs questions méthodologiques importantes se posent alors au statisticien qui dispose de jeux de points à analyser : comment représenter et caractériser spatialement de telles données en utilisant des milliers voire des millions d'observations ? Quels outils statistiques existent et peuvent être mobilisés pour étudier ces observations relatives aux ménages, salariés, firmes, magasins, équipements ou déplacements par exemple ? Comment prendre en compte les caractéristiques qualitatives ou quantitatives des observations étudiées ? Comment mettre en évidence des éventuelles attractions ou répulsions entre les points ou entre différents types de points ? Comment peut-on évaluer la significativité des résultats obtenus ? etc.

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

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

Introduction

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

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

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

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

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

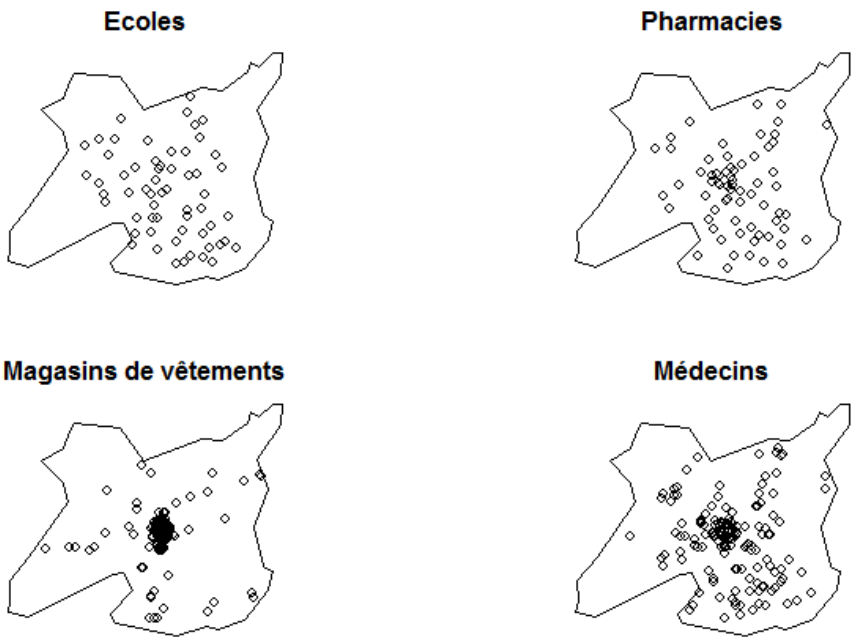


FIGURE 4.1 – Quatre exemples de localisation d'activités dans la commune de Rennes en 2015
Source : *Insee-BPE, calculs des auteurs*

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

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

4.1 Cadre d'analyse : les concepts fondamentaux

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

4.1.1 Configurations et processus

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

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

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

4.1.2 Processus marqués

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

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

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

4.1.3 Fenêtre d'observation

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

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

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

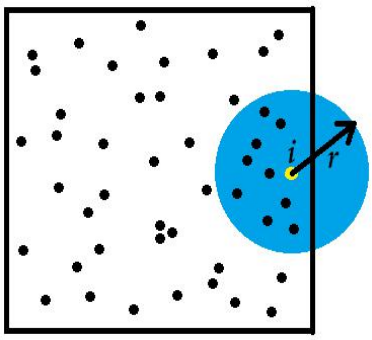


FIGURE 4.2 – Exemple d'effet de bord
Source : *calcul des auteurs.*

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

4.2 Processus ponctuels : une présentation succincte

4.2.1 Le processus de Poisson homogène

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

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

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

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

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

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

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

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

La fonction `runifpoint` du package *spatstat* permet de générer des configurations de points à partir d'un processus binomial uniforme. Par exemple, sur la figure 4.3, 1 000 points sont attendus sur une fenêtre d'observation 10 x 10.

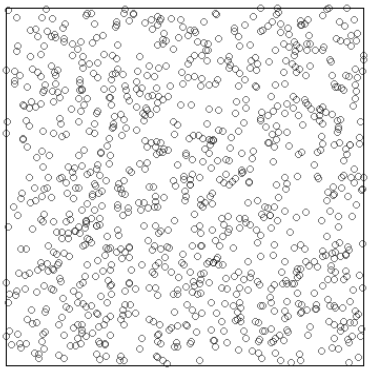


FIGURE 4.3 – Tirage de 1 000 points par un processus binomial uniforme
Source : *package spatstat, calculs des auteurs.*

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

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

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

— **l'indépendance** traduisant le fait que les réalisations dans une région de l'espace n'ont pas d'influence sur les réalisations dans une autre région. Ce n'est pas le cas pour le processus binomial (s'il y a k points dans la région B de W, il y en a $n-k$ dans le complémentaire).

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

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

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

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

Définition 4.2.1 — Processus CSR. Le processus CSR ou processus de Poisson homogène est souvent défini de la façon suivante :

- $P(n(X \cap B) = k) = e^{-\lambda |B|} \frac{\lambda^k |B|^k}{k!}$.
  Cela définit le caractère poissonnien de la distribution **(PP1)** ;
- $\mathbb{E}[n(X \cap B)] = \lambda |B|$.
  Cela définit l'homogénéité **(PP2)** ;
- $n(X \cap B_1)$,..., $n(X \cap B_m)$ sont m variables aléatoires indépendantes **(PP3)** ;
- une fois fixé le nombre de points, la répartition est uniforme **(PP4)**.

Les propriétés **PP2** et **PP3** sont suffisantes pour définir le processus CSR (DIGGLE 1983), et on peut démontrer que les autres en sont les conséquences. D'autres propriétés en découlent. Tout d'abord, la superposition de processus de Poisson indépendants de paramètres λ_1 et λ_2 est un processus de Poisson de paramètre $\lambda_1 + \lambda_2$. Si on élimine des points de façon aléatoire avec une probabilité constante p dans un processus de Poisson (*thinned process* que l'on pourrait traduire par processus amaigri), le processus résultant est toujours un processus de Poisson de paramètre $p\lambda$, où p est le paramètre d'amaigrissement.

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

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

4.2.2 L'intensité, propriété d'ordre 1

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

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

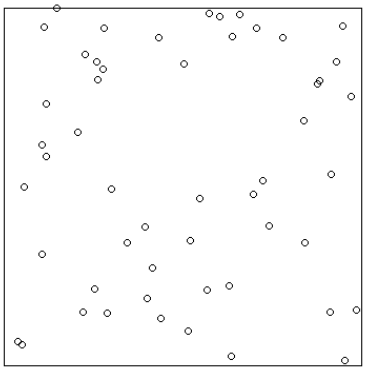


FIGURE 4.4 – Tirage de 50 points par un processus de Poisson homogène
Source : *package spatstat, calculs des auteurs.*

Définition 4.2.2 — Intensité d'un processus. L'intensité est apparue dans la présentation du processus de Poisson où elle était constante (λ). Il existe d'autres processus dans lesquels cette hypothèse est relâchée, et où la fonction d'intensité $\lambda(x)$ est variable. Elle est définie par $\mathbb{E}[n(X \cap B)] = \mu(B) = \int_B \lambda(x)dx$.

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

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

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

4.2.3 Le processus de Poisson inhomogène

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

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

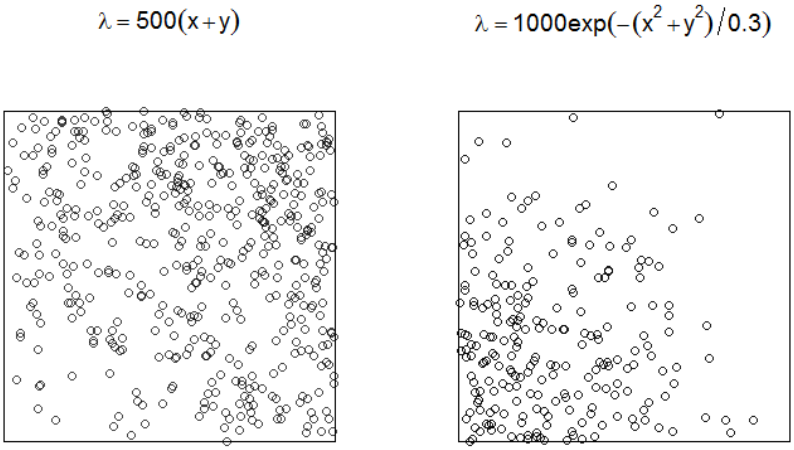


FIGURE 4.5 – Exemples de processus inhomogènes
Source : *package spatstat, calculs des auteurs.*

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

4.2.4 Les propriétés de second ordre

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

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

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

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

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

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

Cette intensité du second ordre peut s'interpréter comme :

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

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

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

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

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

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

4.3 Des processus ponctuels aux répartitions observées de points

4.3.1 Répartition au hasard, agrégation, régularité

Lorsque l'on étudie une distribution de points, deux grandes questions se posent : les points observés sont-ils distribués au hasard ou y a-t-il une interaction ? S'il y a une interdépendance, est-elle de nature agrégative ou répulsive ? Selon les réponses à ces questions, **trois configurations de points** sont généralement mises en évidence : une distribution dite complètement aléatoire, une agrégée et une régulière. Un exemple de ces trois distributions théoriques est représenté sur la figure 4.6. Ces distributions de points sont obtenues à partir de processus ponctuels connus simulés à l'aide du package *spatstat*.

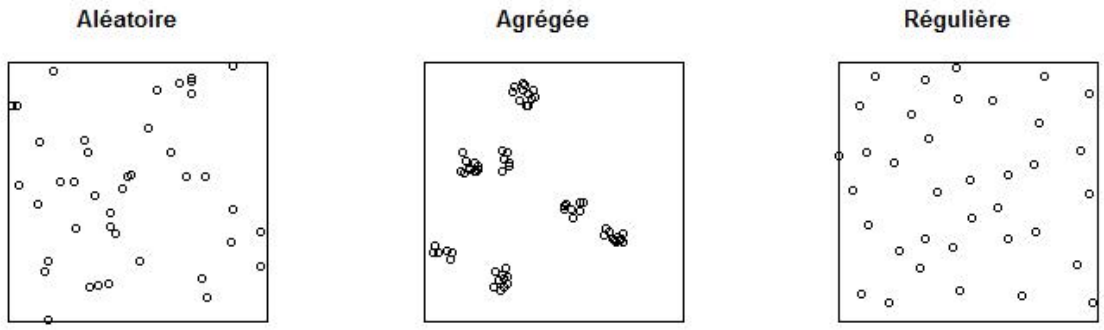


FIGURE 4.6 – Les trois configurations classiques de points
Source : *package spatstat, calculs des auteurs.*

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

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

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

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

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

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

4.3.2 Mises en garde

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

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

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

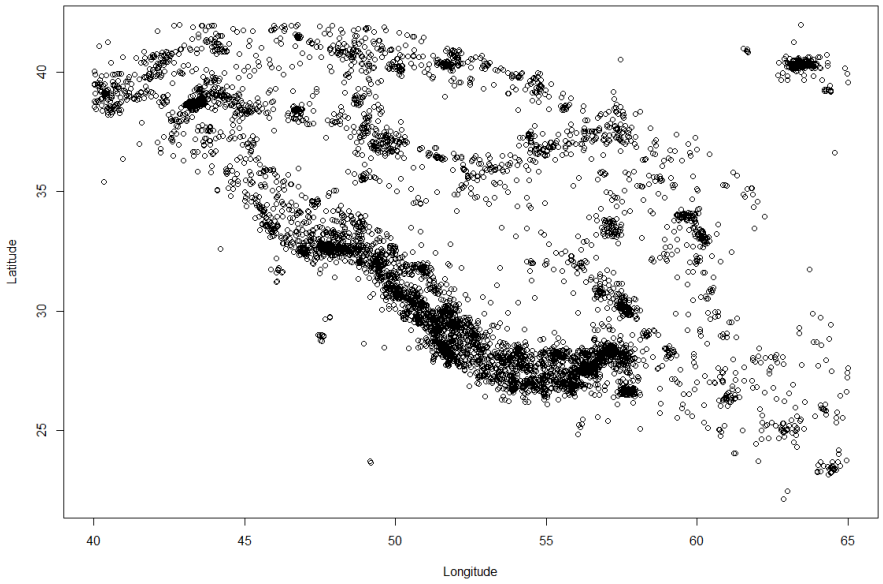


FIGURE 4.7 – Localisation de 5 970 épicentres de séismes en Iran survenus de 1976 à 2016
Source : *package etas, calculs des auteurs.*

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

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

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

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

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

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

4.4.1 La fonction K de Ripley et ses variantes

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

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

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

où n est le nombre total de points sur la fenêtre d'observation, $\mathbf{1}\left\{\left\|x_i - x_j\right\| \leq r\right\}$ est une indicatrice qui vaut 1 si les points i et j sont à une distance au plus égale à r et 0 sinon. $c(x_i, x_j; r)$ correspond à la correction des effets de bord et W à l'aire d'étude.

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

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

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

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

Nous avons rapproché sur la figure 4.8 les trois configurations-types de points vues précédemment et les trois courbes de la fonction K ainsi obtenues. On représente graphiquement en abscisses la distance r et en ordonnées la valeur de la fonction K estimée à cette distance. Avec le package *spatstat*, la fonction K est calculée à l'aide de la fonction `Kest`. Sur la figure 4.8, la fonction K estimée est reportée en noire sur les trois graphiques et la valeur de référence en pointillés rouges. Il vient :

— **lorsque le processus est complètement aléatoire, la courbe s'écarte relativement peu de πr^2.** On peut le constater sur le graphique en bas à gauche de la figure 4.8. La courbe de K reste proche de la valeur de référence πr^2, pour tous les rayons r.

— **dans le cas d'un processus régulier**, on obtient : $\hat{K}(r) < K_{pois}(r)$ puisque si les points se repoussent, ils ont moins de voisins en moyenne dans un rayon r que sous l'hypothèse d'une distribution aléatoire de points. Graphiquement, la courbe K détecte cette répulsion : on constate sur le graphique de droite que la courbe K est située sous la valeur de référence (πr^2) pour tous les rayons.

— **dans le cas d'un processus agrégé**, il y a en moyenne plus de points dans un rayon r autour des points que le nombre attendu sous une distribution aléatoire : par conséquent les points s'attirent et $\hat{K}(r) > K_{pois}(r)$. Graphiquement, la courbe K estimée est cette fois-ci située au dessus de la valeur de référence pour tous les rayons d'étude, comme on peut le noter sur le graphique central reportée sur la figure 4.8.

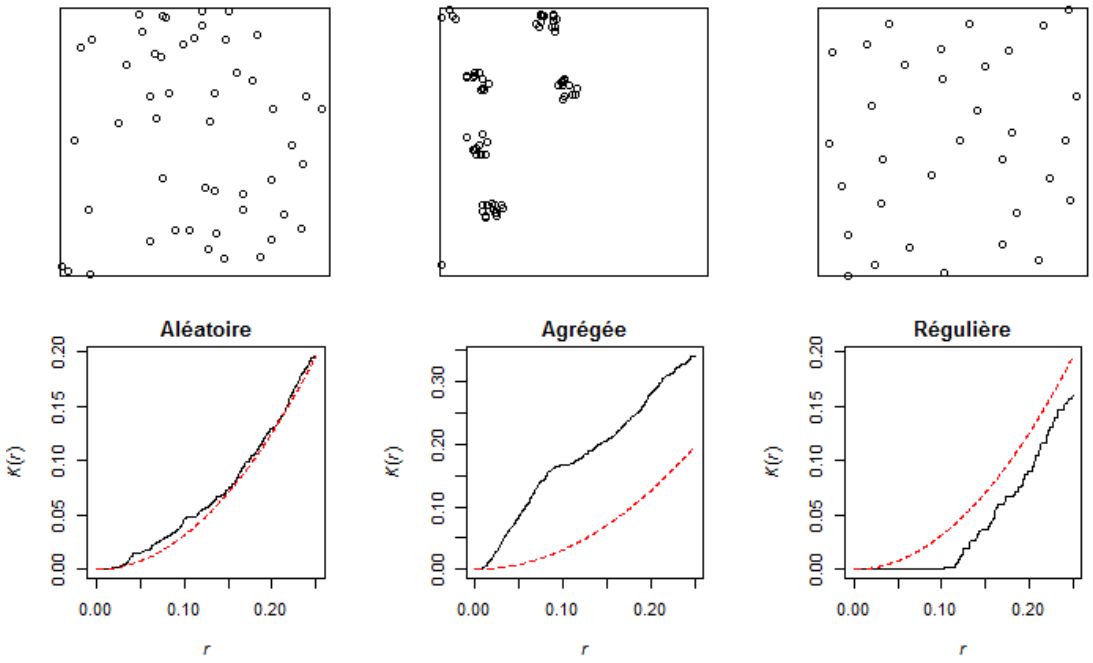


FIGURE 4.8 – Les fonctions K des trois configurations-types
Source : *Package spatstat, calculs des auteurs.*

```
library("spatstat")
par(mfrow=c(2, 3), mar=c(1, 2, 2, 2))
plot(rpoispp(50), main="")
```

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

Soulignons pour terminer quelques points importants.

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

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

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

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

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

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

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

La fonction L de Besag 1977.

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

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

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

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

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

La fonction D de DIGGLE et al. 1991

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

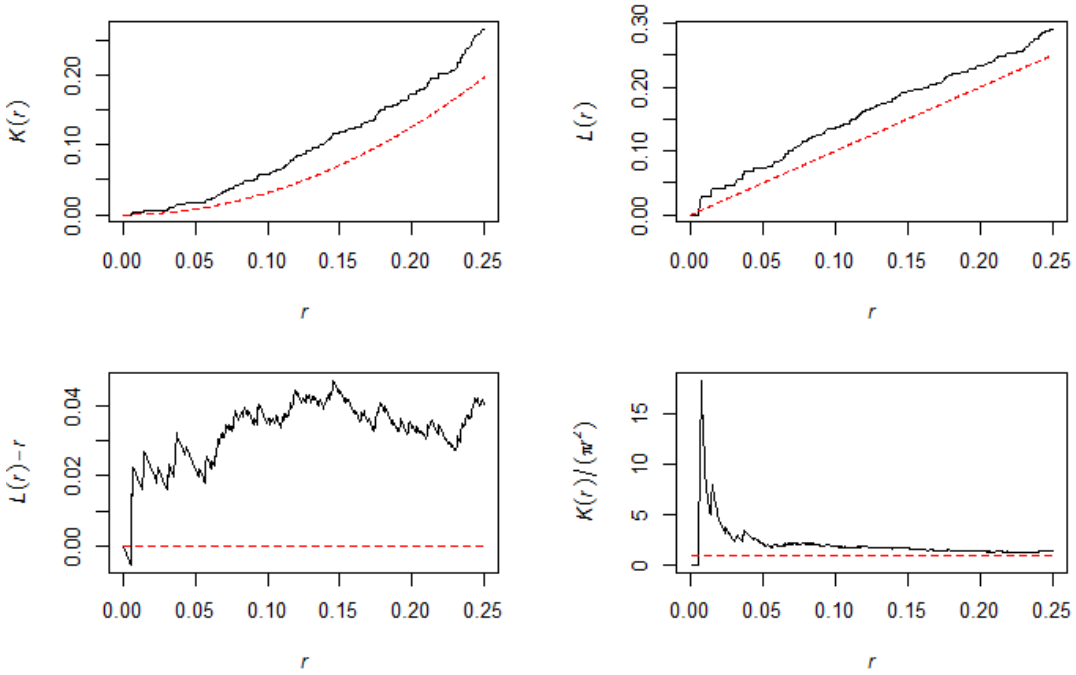


FIGURE 4.9 – Représentation des fonctions K, L, $L-r$ et $K(r)/\pi r^2$ pour l'exemple d'une distribution agrégée
Source : *package spatstat, calculs des auteurs.*

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

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

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

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

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

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

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

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

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

4.4.2 Comment tester la significativité des résultats ?

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

Méthodes de Monte Carlo

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

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

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

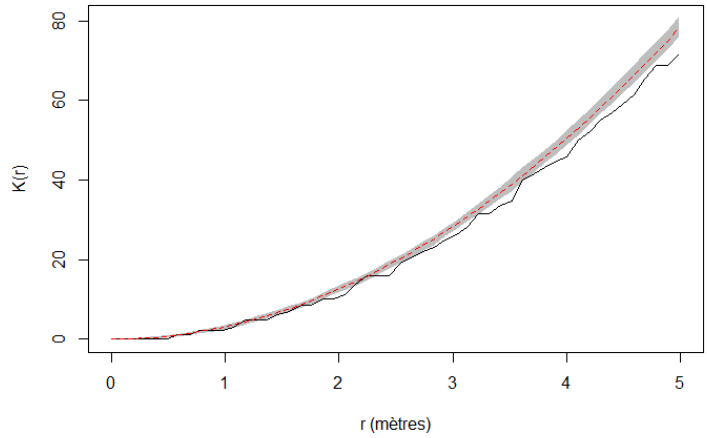


FIGURE 4.10 – Exemple d'enveloppe de confiance pour la fonction K
Source : *package spatstat et dbmss, données "paracou16", calculs des auteurs.*

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

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

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

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

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

Tests analytiques

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

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

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

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

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

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

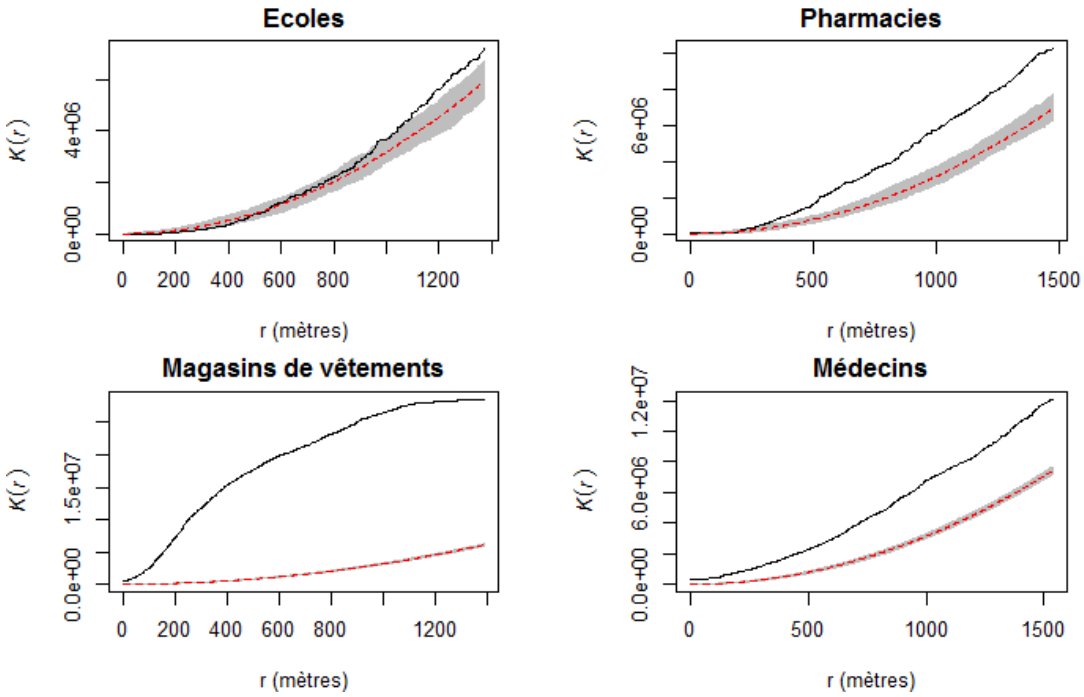


FIGURE 4.11 – Fonctions de Ripley pour les quatre équipements
Source : *Insee-BPE, packages spatstat et dmbss, calculs des auteurs.*

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

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

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

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

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

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

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

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

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

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

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

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

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

— la comparabilité des résultats entre secteurs ;

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

— une référence doit être clairement établie.

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

Quelles pistes de recherches pour des extensions des mesures présentées ?

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

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

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

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

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

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

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

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

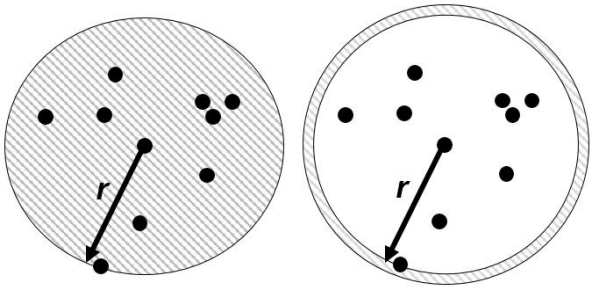


FIGURE 4.12 – Deux notions de voisinages possibles : sur un disque ou sur une couronne
Source : *les auteurs.*

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

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

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

4.5 Mesures fondées sur les distances récemment proposées

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

4.5.1 Indicateur K_d de Duranton et Overman

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

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

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

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

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

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

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

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

4.5.2 Indicateur M de Marcon et Puech

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

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

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

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

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

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

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

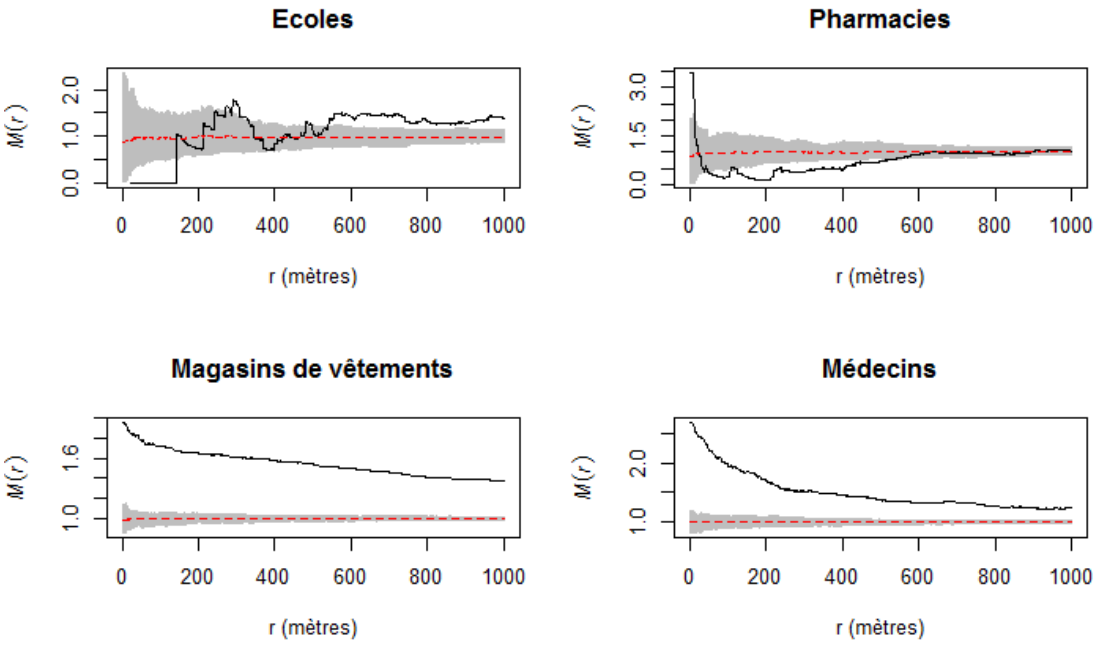


FIGURE 4.13 – Fonctions de Marcon et Puech pour les quatre équipements
Source : *Insee-BPE, packages spatstat et dbmss, calculs des auteurs.*

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

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

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

4.5.3 Autres développements

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

4.6 Processus multitypes

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

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

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

4.6.1 Fonctions d'intensité

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

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

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

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

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

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

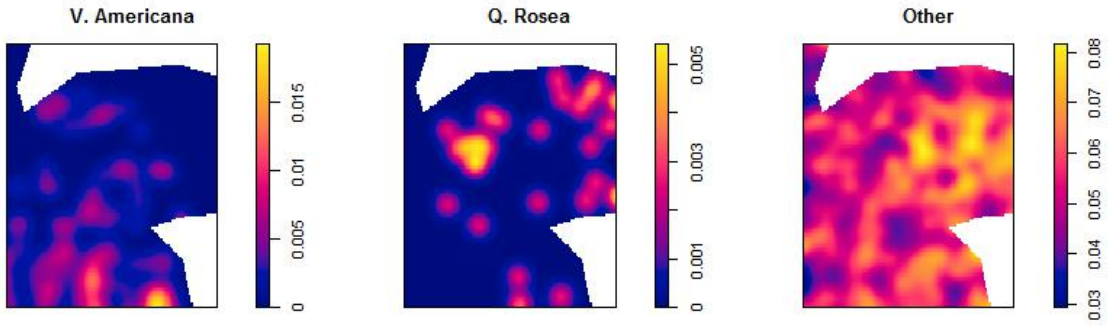


FIGURE 4.15 – Représentation de la densité des arbres d'espèces *Vacapoua americana*, *Qualea rosea* ou autres (mélange) sur le dispositif forestier Paracou16
Source : *données Paracou16 du package dbmss, calculs des auteurs.*

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

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

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

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

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

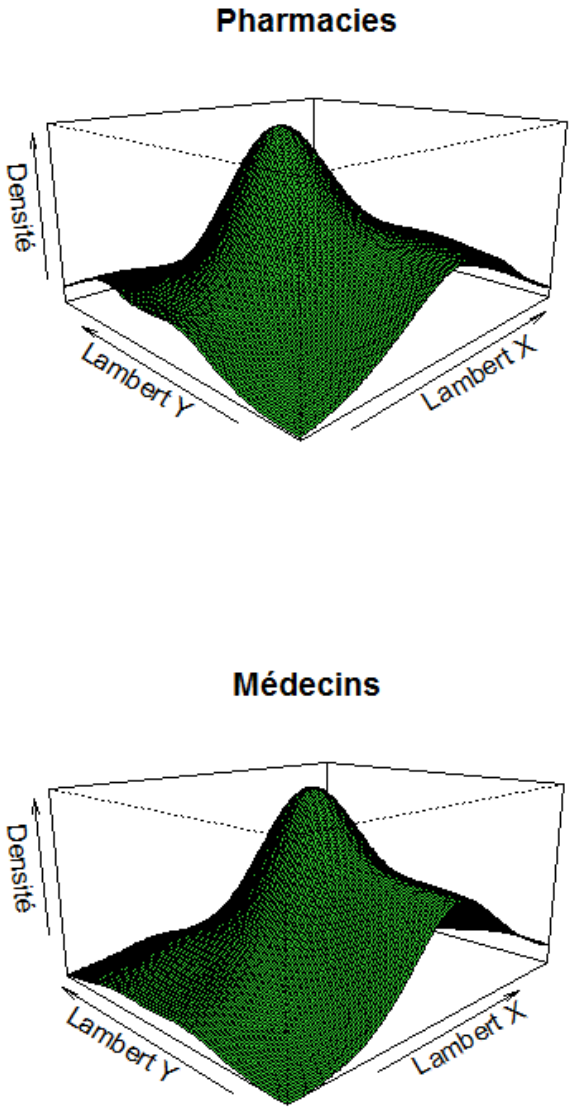


FIGURE 4.16 – Représentation de la densité des pharmacies et des médecins sur Rennes
Source : *Insee-BPE, packages spatstat et dbmss, calculs des auteurs.*

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

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

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

4.6.2 Fonctions intertypes

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

La fonction K intertypes

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

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

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

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

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

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

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

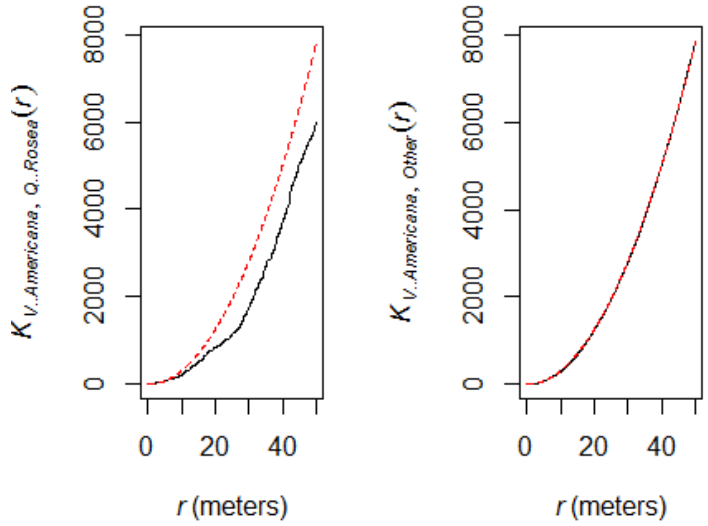


FIGURE 4.17 – Interactions des différentes espèces d'arbres sur le dispositif forestier Paracou16
Source : *données paracou16 du package dbmss, calculs des auteurs.*

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

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

La fonction M intertypes

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

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

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

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

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

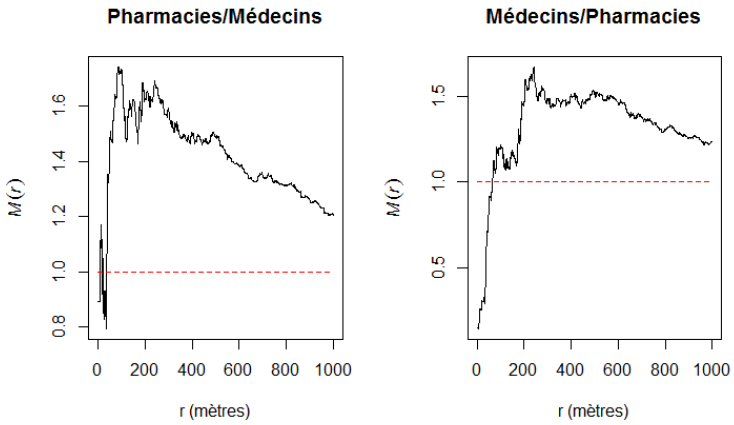


FIGURE 4.18 – Relations de voisinage entre médecins et pharmacies sur Rennes
Source : *packages spatstat et dbmss, calculs des auteurs.*

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

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

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

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

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

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

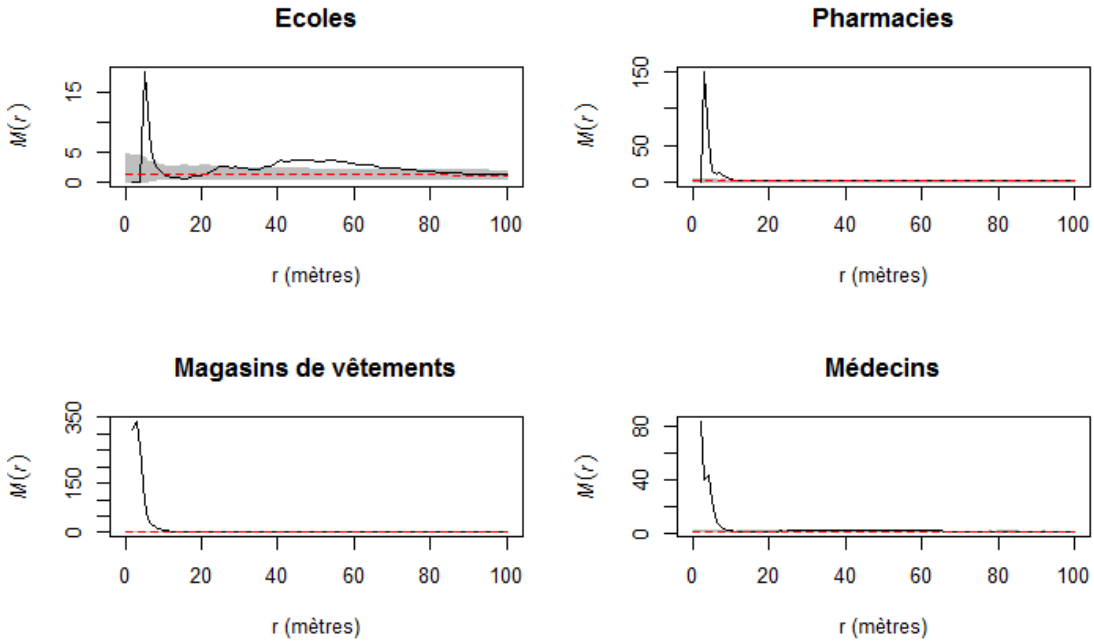


FIGURE 4.19 – Lien équipement/population pour les quatre équipements sur Rennes
Source : *Insee-BPE, package dbmss, calculs des auteurs.*

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

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

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

4.7 Modélisation des processus

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

4.7.1 Cadre général pour la modélisation

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

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

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

4.7.2 Exemples d'application

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

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

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

Conclusion

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

Références - Chapitre 4

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5. Géostatistique

JEAN-MICHEL FLOCH
Insee

Résumé

La géostatistique constitue une branche très importante de la statistique spatiale. Développée à partir de préoccupations très pratiques (recherche minière), elle a fait l'objet, sous l'impulsion de Georges Matheron et de ses collègues de l'école des Mines de Fontainebleau, de très importants développements méthodologiques. Les illustrations les plus simples concernent des problèmes tels que l'interpolation des températures ou des précipitations. Mais les travaux les plus importants concernent les applications géologiques et minières que l'on peut trouver par exemple chez Chilès (CHILES et al. 2009). L'application à des exemples démographiques ou sociaux est plus difficile, mais il semble important de présenter les grandes lignes de la méthode : traitement de la stationnarité avec l'introduction de la stationnarité intrinsèque ; introduction du semi-variogramme pour l'étude des relations spatiales ; interpolation des données par la méthode du krigeage. Au delà des applications minières, l'analyse variographique peut être utilisée dans des modèles mixtes, pour analyser les résidus.

R La lecture préalable du chapitre 1 : "Analyse spatiale descriptive" et du chapitre 4 : "Les configurations de points" est recommandée.

La modélisation des données spatiales est rendue difficile du fait que l'on n'observe qu'une seule réalisation du phénomène. Dans le cas des données ponctuelles, on n'observe également qu'une seule réalisation, pour laquelle on dispose de toutes les données. Pour les données continues (potentiellement observables en tout point de l'espace), on ne dispose que de données partielles, à partir desquelles on peut être amené à prédire des valeurs aux points non observés. C'est ce manque d'information qui va conduire à l'utilisation de modèles probabilistes.

L'aléatoire n'est pas une propriété du phénomène, mais une caractéristique du modèle utilisé pour le décrire. La géostatistique, qui étudie les phénomènes continus, a permis le développement de méthodes spécifiques pour étudier les relations spatiales entre les observations et construire des outils prédictifs.

La géostatistique doit son nom aux origines minières de la discipline (Krige, Matheron). Nombre des concepts fondamentaux de la discipline sont issus des travaux de Georges Matheron (*variable régionalisée, fonction aléatoire, hypothèse intrinsèque, effet de pépite* [1], MATHERON et al. 1965). La figure 5.1 présente un résumé illustrant le cheminement de la réalité vers un modèle plus abstrait, modèle qui permettra lui-même d'agir d'une façon que l'on espère optimale (CHAUVET 2008).

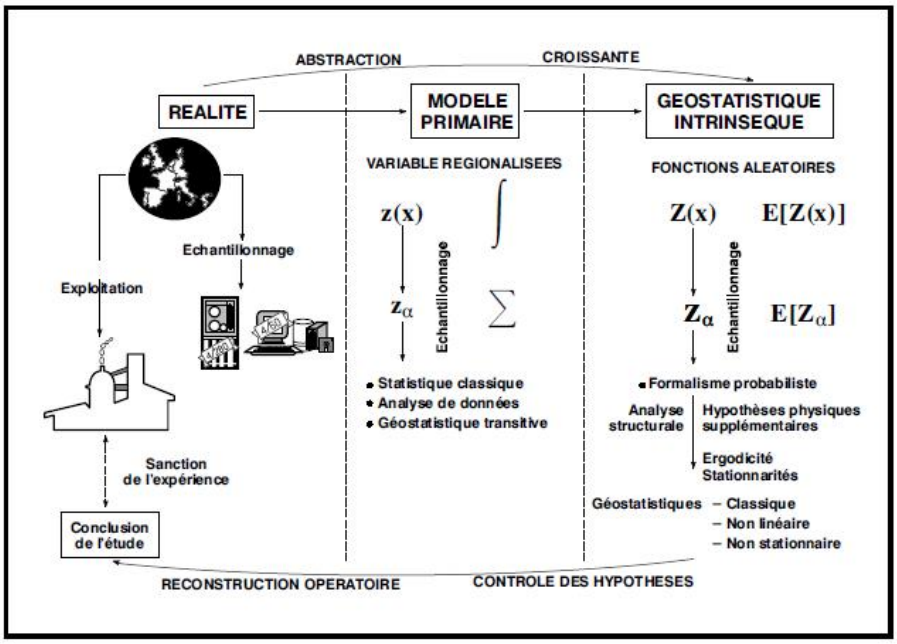


FIGURE 5.1 – Schéma d'une analyse géostatistique
Source : CHAUVET *2008*

5.1 Fonctions aléatoires

5.1.1 Définitions

Comme dans la figure 5.1, on note $z(s)$ la *variable régionalisée*, et $Z(s)$ la *fonction aléatoire*, la lettre s désignant la position dans l'espace. On conservera ici cette formulation spécifique à la

1. Ces termes sont définis plus loin.

géostatistique. Un phénomène qui se déploie dans l'espace est qualifié de régionalisé. Une variable régionalisée est une fonction qui décrit de manière satisfaisante ce phénomène. C'est un premier niveau d'abstraction, où l'on reste dans la description, sans recourir à un modèle probabiliste. Si on ne fait pas d'hypothèse supplémentaire, on reste dans le cadre de la *géostatistique transitive*.

L'étape ultérieure, qualifiée de *géostatistique intrinsèque*, introduit la notion de fonction aléatoire. Elle résulte d'un choix, celui de considérer la variable régionalisée comme la réalisation d'une fonction aléatoire. Ce choix permet d'utiliser des outils probabilistes puissants, la contrepartie étant un éloignement de la réalité. Le modèle probabiliste est un intermédiaire de calcul dont on attend qu'il rende des services pour la compréhension du phénomène régionalisé.

La fonction aléatoire est caractérisée complètement par la donnée de sa fonction de répartition

$$F(s_1, s_2, ...s_n; z_1, z_2, ...z_n) = P\{Z(s_1) \leq z_1, Z(s_2) \leq z_2, ..., Z(s_n) \leq z_n\}. \tag{5.1}$$

Comme on ne dispose que d'une seule réalisation de notre phénomène régionalisé, il faut trouver une autre façon de faire de l'inférence. Citant MATHERON et al. 1965 : " Pour que l'inférence soit possible, il est nécessaire d'introduire des hypothèses supplémentaires sur la fonction aléatoire $Z(s)$ de façon à réduire le nombre des paramètres dont dépend sa loi. Tel est le but de l'hypothèse stationnaire que nous allons définir : une fonction stationnaire se répète en quelque sorte elle-même dans l'espace, et cette répétition rend à nouveau possible l'inférence statistique à partir d'une réalisation unique." On va de fait traiter chaque observation comme la réalisation d'une variable aléatoire.

5.1.2 Stationnarité

Trois acceptions de la stationnarité sont utilisées en géostatistique :
- la stationnarité stricte ;
- la stationnarité au second ordre ;
- la stationnarité intrinsèque.

Définition 5.1.1 — Stationnarité stricte. La stationnarité stricte renvoie directement à la loi de probabilité du processus. Il y a stationnarité stricte si en se déplaçant par translation, toutes les caractéristiques de la fonction aléatoire restent les mêmes.
Formellement la distribution jointe des $Z(s_i)$ est la même que celle des $Z(s_i+h)$, h indiquant une translation par rapport à la position initiale. Cette forme de stationnarité n'est pas opérationnelle, et très restrictive.

Définition 5.1.2 — Stationnarité du second ordre. La stationnarité au second ordre ou stationnarité faible n'impose plus de conditions sur la loi de probabilité, mais seulement sur la moyenne et la covariance. Ces indicateurs doivent être invariants par translation.
Sachant que l'on décompose $Z(s)$ en une composante déterministe et une composante aléatoire $Z(s) = m(s) + R(s)$

la stationnarité au second ordre requiert les conditions suivantes :
- $\mathbb{E}[Z(s)] = m(s) \forall s$.
  L'invariance de l'espérance par translation entraîne la constance de la composante déterministe.
  $m(s+h) = m(s) = m \forall s$;
- La variance est constante : $\mathbb{E}\left[(Z(s)-m)^2\right] = \sigma^2$;
- La covariance ne dépend que du décalage spatial :
  $Cov[Z(s+h), Z(s)] = \mathbb{E}[(Z(s+h)-m)(Z(s)-m)] = C(h)$.

En pratique, cette hypothèse de stationnarité s'avère souvent trop forte. La limite la plus importante vient de ce que la moyenne peut changer sur le territoire d'intérêt, et que la variance peut ne pas être bornée lorsque cette aire d'intérêt croit. C'est Georges Matheron qui a tiré les conséquences de ces limites de la stationnarité faible en proposant la notion, encore plus faible, de la stationnarité intrinsèque (MATHERON et al. 1965).

Définition 5.1.3 — Stationnarité intrinsèque. L' hypothèse de la stationnarité intrinsèque est la suivante : $\mathbb{E}\left[(Z(s+h)-Z(s))^2\right]=0$.
Les accroissements peuvent être stationnaires sans que le processus lui-même le soit.

On peut alors définir une nouvelle fonction, appelée *variogramme*, fondée sur les différences entre valeurs et valeurs décalées, et qui ne dépend que du décalage :

$$\gamma(h)=\frac{1}{2}\mathbb{E}\left[Z(s+h)-Z(s)\right]^2 \tag{5.2}$$

La stationnarité à l'ordre 2 entraîne la stationnarité intrinsèque, mais l'inverse n'est pas vrai. Une fonction aléatoire peut permettre le calcul d'un variogramme sans qu'il en soit de même pour la covariance et la fonction d'autocorrélation.

5.2 Variabilité spatiale

5.2.1 Covariance et corrélogramme

Définition 5.2.1 — Covariance. La fonction de covariance va permettre de prendre en compte les relations entre l'ensemble des paires de points. Si on prend en compte deux points s_i et s_j, la covariance peut être définie par l'équation 5.3.

$$Cov\left[Z(s_i),Z(s_j)\right]=\mathbb{E}\left[(Z(s_i)-m)\,(Z(s_j)-m)\right] \tag{5.3}$$

Lorsque le processus est stationnaire au second ordre, la covariance ne va plus dépendre que de la distance entre les points $|s_i-s_j|$. Si on note h cette distance, on va définir $C(h)$ calculée pour toutes les valeurs de h en prenant en compte tous les couples de points situés à une distance h les uns des autres. Cette fonction de covariance $C(h)$ est définie par l'équation 5.4.

$$C(h)=Cov\left[Z(s+h),Z(s)\right]=\mathbb{E}\left[(Z(s+h)-m)\,(Z(s)-m)\right] \tag{5.4}$$

Elle traduit la façon dont évoluent la covariance des observations lorsque leur distance augmente. Lorsque h est égal à 0, la covariance est égale à la variance.

$$C(0)=\mathbb{E}\left[(Z(s)-m)^2\right]=\sigma^2 \tag{5.5}$$

Les propriétés de la fonction de covariance sont les suivantes :

$$C(-h)=C(h) \tag{5.6}$$
$$|C(h)|\leq C(0). \tag{5.7}$$

Pour que la fonction de covariance soit dite *admissible*, il faut que la variance d'une combinaison linéaire de variables soit positive : $Var\left[\sum_{i=1}^{n}\lambda_i z(s_i)\right]=\sum_{i=1}^{n}\sum_{j=1}^{n}C(s_i-sj)$.

Cela entraîne que C soit *semi-définie positive*.

Définition 5.2.2 — Fonction d'autocorrélation. On définit la fonction d'autocorrélation $\rho(h)$ comme une fonction de h par le rapport $\frac{C(h)}{C(0)}$. Sa valeur est comprise entre -1 et +1. On peut montrer les relations suivantes lorsque la stationnarité à l'ordre 2 est vérifiée :

$$\begin{aligned} \gamma(h) &= C(0) - C(h) \\ \gamma(h) &= \sigma^2 (1 - \rho(h)) . \end{aligned} \tag{5.8}$$

Encadré 5.2.1 — Estimation de la fonction de covariance. La fonction de covariance est estimée à partir des $n(h)$ paires de points, comme définies ci dessous, pour i variant de 1 à $n(h)$.

$$\widehat{C}(h) = \frac{1}{n(h)} \sum_{i=1}^{n(h)} (z(s_i) - m)(z(s_i + h) - m) \tag{5.9}$$

avec $n(h) = Card\left\{(s_i, s_j) / \left|s_i - s_j\right| \approx h\right\}$

5.2.2 Variogramme

On trouve dans la littérature les expressions de *variogramme* ou de *semi-variogramme*. Certains auteurs (MATHERON et al. 1965) estiment qu'il faut utiliser le terme de semi-variogramme pour $\gamma(h)$ tel que défini dans l'équation 5.10, le variogramme correspondant à $2\gamma(h)$. C'est le choix que nous faisons dans cet article.

Des trois indicateurs que sont la fonction de covariance, la fonction d'autocorrélation et le variogramme, ce dernier est le plus utilisé dans la mesure où il renvoie à la forme la plus faible de stationnarité et donc aux conditions les moins restrictives sur le comportement local de la moyenne.

$$\gamma(h) = \sigma^2 (1 - \rho(h)) \tag{5.10}$$

Les propriétes du variogramme sont les suivantes :

$$\begin{aligned} &\gamma(h) = \gamma(-h) \\ &\gamma(0) = 0 \\ &\frac{\gamma(h)}{\|h\|^2} \to 0 \quad quand \quad \|h\| \to \infty \end{aligned} \tag{5.11}$$

Pour tout ensemble de réels $\{a_1, a_2, ...a_m\}$ vérifiant $\sum_{i=1}^{m} a_i = 0$, on a la propriété suivante :

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j \gamma(s_i - s_j) \leq 0 \tag{5.12}$$

Lorsque le processus est isotrope :

$$\gamma(h) = \gamma(\|h\|) \tag{5.13}$$

Encadré 5.2.2 — Estimation du variogramme expérimental. Un variogramme expérimental peut être estimé à partir des couples de points définis comme précédemment.

$$\widehat{\gamma}(h) = \frac{1}{n(h)} \sum_{i=1}^{n(h)} (z(s_i + h) - z(s_i))^2 \tag{5.14}$$

avec $n(h) = Card\left\{(s_i, s_j) / \left|s_i - s_j\right| \approx h\right\}$
Le variogramme peut être estimé selon différentes directions pour mettre en évidence l'éventuelle anisotropie du phénomène étudié.

5.2.3 Application empirique

La géostatistique avec R

Les packages dédiés à la géostatistique dans ce chapitre sont **gstat** et **geoR**, les plus couramment utilisés. Sur le site du CRAN, de nombreux autres packages sont proposés. On trouvera ci dessous une liste commentée réalisée par Roger Bivand.

Le package **gstat** fournit un large éventail de fonctions pour la géostatistique univariée et multivariée, y compris pour les ensembles de données plus volumineux, tandis que **geoR** et **geoRglm** contiennent des fonctions pour la géostatistique basée sur un modèle. Le diagnostic du variogramme peut être effectué avec **vardiag**. L'interpolation automatisée utilisant gstat est disponible dans **automap**. Cette famille de packages est complétée par **intamap** avec des procédures d'interpolation automatique et **psgp**, qui implémente le krigeage de processus gaussien clairsemé. Un large éventail de fonctions similaires se trouve dans le package **fields**. Le package spatial est livré avec la base R et contient plusieurs fonctions principales. Le package **spBayes** est compatible avec les modèles univariés et multivariés gaussiens avec MCMC. **rampes** est un autre ensemble de modélisation géostatistique bayésiennne. Le package **geospt** contient des fonctions de base géostatistiques et radiales, y compris la prédiction et la validation croisée. En outre, il inclut des fonctions pour la conception de réseaux d'échantillonnage spatial optimaux basés sur la modélisation géostatistique. spsann est un autre package qui offre des fonctions pour optimiser les configurations d'échantillons, en utilisant un recuit spatial simulé. Le package **geostatsp** offre des fonctions de modélisation géostatistique utilisant des objets Raster et SpatialPoints. Les modèles non gaussiens sont adaptés à l'aide de l'INLA, et les modèles géostatistiques gaussiens utilisent l'estimation du maximum de vraisemblance. Le package **FRK** est un outil de modélisation et de prédiction spatiales / spatio-temporelles avec de grands ensembles de données. L'approche, discutée dans Cressie et Johannesson (2008), décompose le champ, et donc la fonction de covariance, en utilisant un ensemble fixe de n fonctions de base, où n est généralement beaucoup plus petit que le nombre de points de données (ou polygones).

Le package RGeostats

RGeostats est un package du langage R. Il a été développé par l'équipe de géostatistique du centre de Géosciences de des Mines ParisTech. Il met en œuvre toutes les fonctions géostatistiques disponibles de la librairie (commerciale) Geoslib (écrite en C/C++). Il bénéficie donc de l'expérience accumulée dans le domaine minier, et il est spécialement dédié à de telles applications. Il permet de réaliser toutes les implémentations présentées dans ce chapitre, et de traiter les effets de support.

RGeostats permet à l'utilisateur de R, moyennant son chargement et installation, d'accéder aux fonctionnalités usuelles de la Géostatistique. C'est également une plate-forme qui permet à l'équipe de Géostatistique du Centre de Géosciences de développer des prototypes pour l'application de nouveaux modèles (ex : simulations booléen, simulations bi-plurigaussiennes) ou nouvelles techniques (écoulement de fluides, simulations du temps de première arrivée en géophysique. Les fonctions sont décrites, mais le code n'est pas accessible, et il y a peu d'exemples.

RGeostats est disponible par simple téléchargement à l'adresse suivante : cg.ensmp.fr/rgeostats.

Analyses exploratoires

L'application proposée utilise les données *Swiss rainfall* (précipitations en Suisse). Cette base de données est très utilisée dans les études spatiales, notamment dans DIGGLE et al. 2003. Elle est fournie dans le package *geoR* de R. Les observations correspondent aux relevés pluviométriques de 467 stations météorologiques de Suisse, et sont effectuées le 8 mai 1986. Il s'agit bien d'une donnée continue, les précipitations pouvant être potentiellement relevées en tout point du territoire. Elle peut donc relever d'une modélisation géostatistique, mais elle est plus simple à appréhender que les données minières. En plus de la pluviométrie, mesurée en millimètres, on trouve des données sur l'altitude des stations météorologiques. Dans le package *geoR*, on dispose de trois bases Sic (*Spatial interpolation comparaison*) :

— *Sic.100* : échantillon de 100 observations qui pourront servir à effectuer les interpolations ;
— *Sic.367* : observations non incluses dans l'échantillon qui permettront de comparer les estimations et les observations ;
— *Sic.all* : ensemble.

On utilisera ici les fonctionnalités de *geoR*, mais *gstat* et *RGeostats* fournissent des outils d'analyse.

On trouvera dans la figure 5.2, les données échantillonnées (en vert) et les données témoins, les cercles étant proportionnels à la pluviométrie enregistrée.

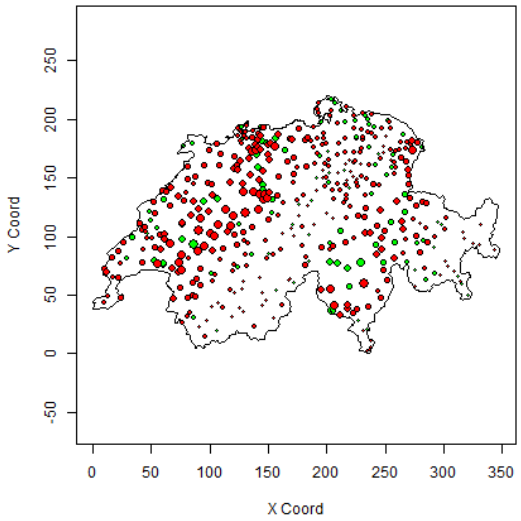


FIGURE 5.2 – Les précipitations en Suisse
Source : *Swiss rainfall du package geoR*

```
library(geoR)
points(sic.100, borders=sic.borders,col="green")
points(sic.367, borders=sic.borders,col="red",add=TRUE)
```

Le package *geoR* fournit quelques représentations descriptives grâce à la fonction `plotgeodata`. On trouve dans la figure 5.3, de gauche à droite et de haut en bas :

— la représentation du niveau de la pluviométrie, selon les quantiles de la variable ;
— la pluviométrie en fonction de la latitude ;
— la pluviométrie en fonction de la longitude ;
— l'histogramme des données de pluviométrie.

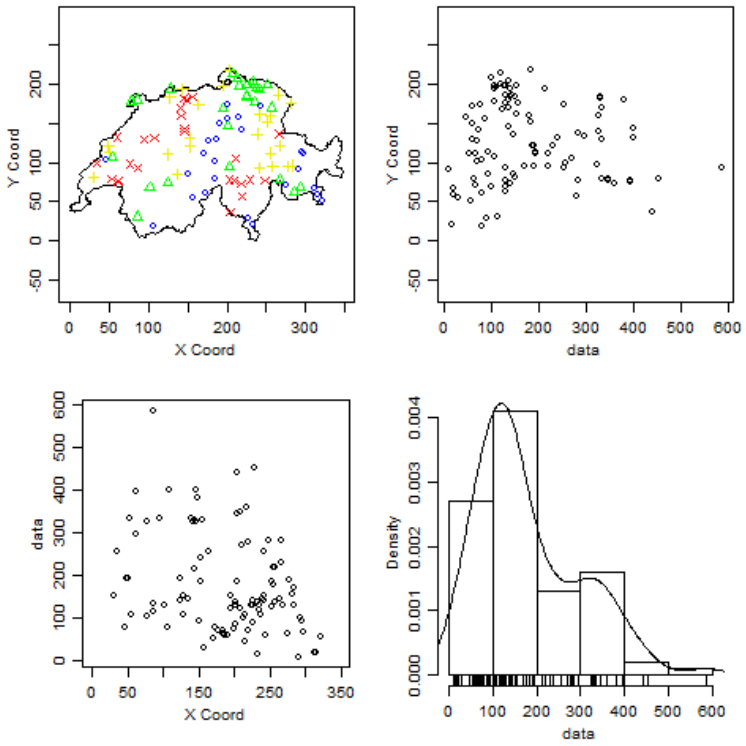


FIGURE 5.3 – Quelques statistiques descriptives
Source : *Swiss rainfall du package geoR*

```
library(geoR)
plot.geodata(sic.100,bor=sic.borders)
```

L'histogramme de la figure 5.3 laisse penser que la distribution de la variable n'est pas gaussienne, et qu'une transformation des données pourrait être envisagée puisque les méthodes les plus courantes n'ont des propriétés intéressantes que dans le cadre gaussien.

Nuée variographique et variogramme expérimental

En géostatistique intrinsèque, la **nuée variographique** est un nuage de points des données exprimant leur variabilité selon leurs interdistances. La nuée variographique fournit la représentation graphique des valeurs intervenant dans le calcul du variogramme. Pour un jeu de données de la variable Z aux points $(s_1,.,s_i,.,s_n)$, il représente les points d'abscisses $\|s_i - s_j\|$ et d'ordonnées $\frac{1}{2}[z(s_i) - z(s_j)]^2$. On peut la représenter sous forme de nuage de points et sous forme de boxplot (voir figure 5.4).

```
library(geoR)
vario.b<-  variog(sic.100,option =c ("bin", "cloud", "smooth"),
bin.cloud=TRUE)
vario.c <- variog(sic.100, op="cloud")
bplot.xy(vario.c$u,vario.c$v, breaks=vario.b$u,col="grey80", lwd=2,cex=0.1,
    outline=FALSE)
```

Ces représentations étant peu lisibles, la représentation la plus utile est le variogramme expérimental (défini en encadré 5.2.2), présenté en figure 5.6 à partir d'un schéma de construction présenté en figure 5.5.

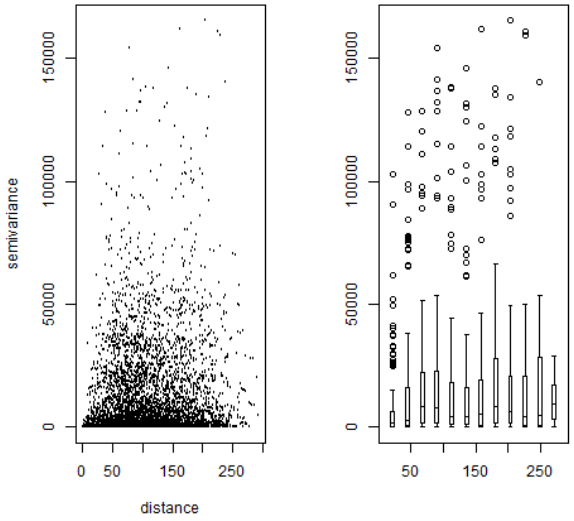


FIGURE 5.4 – Nuée variographique
Source : *Swiss rainfall du package geoR*

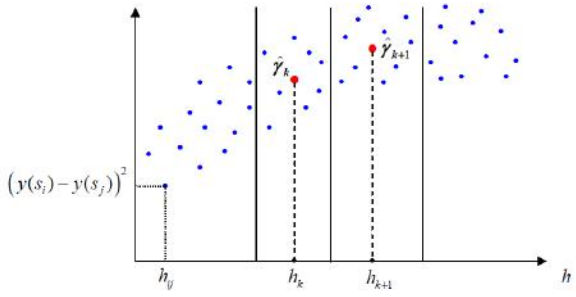


FIGURE 5.5 – Variogramme expérimental : schéma de construction
Source : *Swiss rainfall du package geoR*

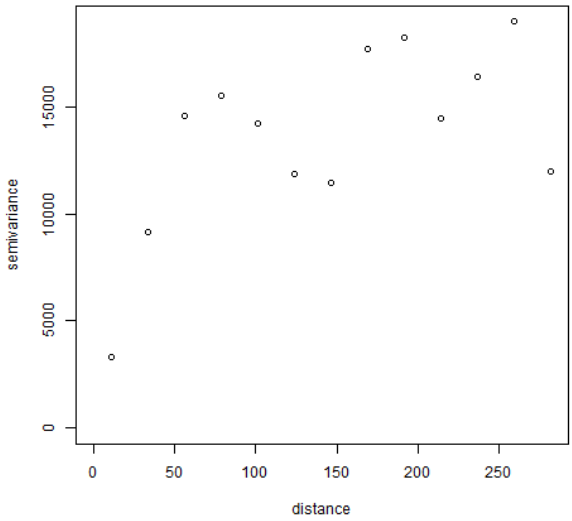


FIGURE 5.6 – Variogramme expérimental
Source : *Swiss rainfall du package geoR*

```
library(geoR)
vario.ex<-  variog(sic.100, bin.cloud=TRUE)
plot(vario.ex)
```

Dans la figure 5.6, tous les points observés sont pris en compte pour le calcul du variogramme. Mais les phénomènes étudiés ne sont pas forcément isotropes, et il peut être utile de calculer des variogrammes selon plusieurs directions de l'espace (voir figure 5.7).

```
library(geoR)
vario4<-variog4(sic.100)
plot(vario4,same=FALSE)
```

5.3 Ajustement du variogramme

On verra dans la partie 5.4 consacrée au krigeage que la valeur des estimateurs dépend des observations et de la structure d'autocorrélation spatiale, appréhendée par le variogramme. L'analyse variographique n'est donc pas qu'un point de passage. Elle constitue le point central de la démarche géostatistique. Les variogrammes empiriques présentés dans la partie 5.2 ne sont pas directement utilisables car ils ne respectent pas les propriétes énoncées dans la partie 5.2.2. Pour être utilisés dans les modèles géostatistiques, il faut au préalable les ajuster à des modèles théoriques ayant des formes analytiques bien définies, ce qui implique une vision de ce que doit être un semi-variogramme.

5.3.1 Allure générale du variogramme

On commencera par présenter le modèle le plus classique, à partir duquel seront construits les variogrammes théoriques qui permettront de définir entre autres les équations de krigeage (voir la partie 5.4).

Ce variogramme a une forme qui est d'abord croissante, jusqu'à un certain palier. La valeur de h correspondant à ce palier est appelée la *portée*. On le comprend en se référant à la relation entre

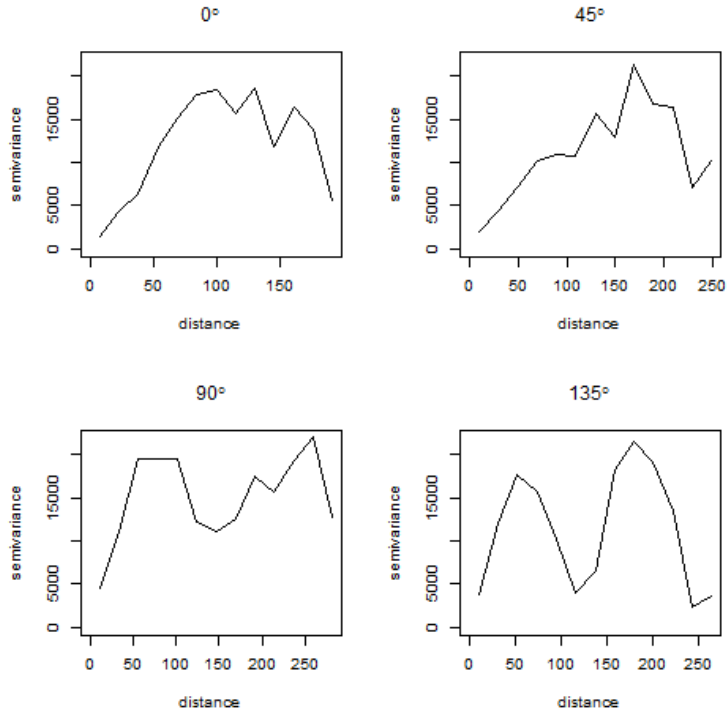


FIGURE 5.7 – Variogrammes directionnels
Source : *Swiss rainfall du package geoR*

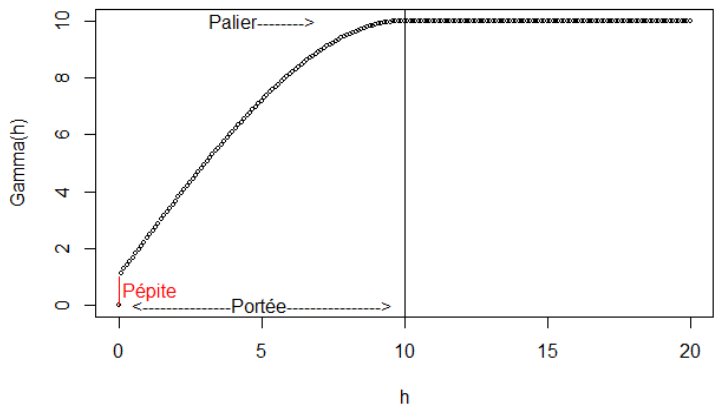


FIGURE 5.8 – Variogramme théorique

la covariance, lorsque celle-ci est définie et le semi-variogramme, à savoir $\gamma(h) = C(0) - C(h)$. La covariance est très fréquemment une fonction décroissante de la distance, ce qui implique la croissance du semi-variogramme, mais ce n'est pas toujours le cas (modèle sinus-cardinal par exemple).

À une certaine distance, qui correspond à la portée, la covariance va s'annuler. Cette portée est la portée de la dépendance spatiale. Il n'y a plus de relation entre les valeurs observées à une distance observée au-delà de cette portée. Pour le semi-variogramme cela signifie qu'au-delà de cette portée, sa valeur est constante, et que l'on a : $C = C(0) = \sigma^2$.

Pour $h = 0$, la valeur du variogramme est par définition nulle. Mais on constate en pratique que pour des valeurs très proches de 0, le variogramme prend des valeurs supérieures à 0 et qu'il y a de fait une discontinuité à l'origine. On appellera *pépite* la limite du variogramme en zéro. Comme l'explique Matheron (MATHERON et al. 1965) : "La notion d'échelle joue ici un rôle primordial. À l'échelle de la dizaine de mètres, un phénomène de transition dont la portée est centimétrique ne se manifeste sur $\gamma(h)$ que comme une discontinuité à l'origine, c'est à dire un effet de pépite." Elle représente la variation entre deux mesures effectuées à des emplacements infiniment proches, et peut donc provenir de deux effets :

— une variabilité de l'instrument de mesure : la pépite mesure donc en partie l'erreur statistique de l'instrument de mesure
— un réel effet pépite : une variation brutale du paramètre mesuré ; le cas historique est le passage sans transition d'une pépite d'or à un sol ne contenant quasiment pas d'or.

On peut rencontrer d'autres formes de variogramme. La figure 5.9 montre deux cas classiques. Le premier est celui du **variogramme linéaire**, le second celui du **pur effet de pépite**. Lorsque le variogramme n'est pas borné, la moyenne et la variance ne sont pas définies : cela indique le plus souvent une tendance à grande échelle, qu'il sera nécessaire de modéliser. L'effet de pépite pur traduit l'absence de dépendance spatiale.

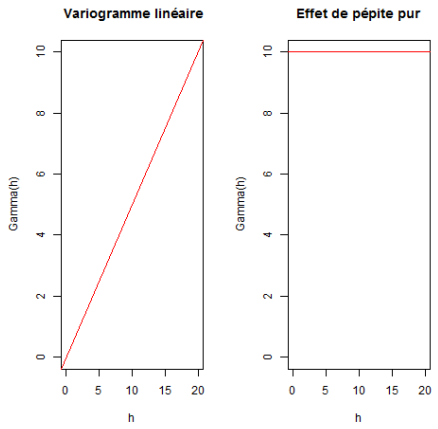


FIGURE 5.9 – Deux semi-variogrammes atypiques

5.3.2 Les variogrammes usuels

La littérature géostatistique propose de nombreuses fonctions qui satisfont les propriétés du semi-variogramme tel qu'il est présenté dans la figure 5.8. Ces fonctions paramétrées doivent permettre de décrire les différentes composantes (portée, palier, pépite). Elles doivent aussi gérer le comportement de la fonction à l'origine (tendance linéaire, tangence horizontale ou verticale).

On ne présentera ici que quatre exemples de modèles de variogramme (figure 5.10), les autres

étant décrits dans les ouvrages de référence (ARMSTRONG 1998, CHILES et al. 2009,WALLER et al. 2004).

Définition 5.3.1 — Le modèle sphérique.

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + c_s \left[\frac{3}{2} \left(\frac{h}{a} \right) - \frac{1}{2} \left(\frac{h}{a} \right)^3 \right] & 0 < h \leq a \\ c_0 + c_s & h > a \end{cases} \tag{5.15}$$

Définition 5.3.2 — Le modèle exponentiel.

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + c_s \left[1 - exp\left(-\frac{h}{a} \right) \right] & h > 0 \end{cases} \tag{5.16}$$

Définition 5.3.3 — Le modèle gaussien.

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + c_s \left[1 - exp\left(-(\frac{h}{a})^2 \right) \right] & h > 0 \end{cases} \tag{5.17}$$

Définition 5.3.4 — Le modèle puissance.

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + bh^p & h > 0 \end{cases} \tag{5.18}$$

Définition 5.3.5 — Modèle de Matern.

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_s \left[1 - \frac{\frac{h}{a}}{2^{\alpha-1}\Gamma(\alpha)} K_\alpha \left(\frac{h}{a} \right) \right] & h > 0 \end{cases} \tag{5.19}$$

où Γ désigne la fonction gamma et K_α la fonction de Bessel modifiée de seconde espèce de paramètre α.

Définition 5.3.6 — Modèle sinus cardinal.

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_s \left[1 - \frac{a}{h} sin\left(\frac{h}{a} \right) \right] & h > 0 \end{cases} \tag{5.20}$$

Comme l'expliquent Matheron et les géostatisticiens de l'école des Mines de Fontainebleau, la modélisation est une affaire de choix. Le choix du modèle théorique est un moment décisif dans la démarche du géostatisticien, mais on ne peut pas associer *a priori* un variogramme théorique à tel ou tel type de processus. Il faut tenir compte à la fois de la connaissance empirique du phénomène, de l'allure du variogramme expérimental obtenu. S'agissant de l'effet de pépite, on peut par exemple penser qu'il n'est pas adapté à des données comme les précipitations.

5.3.3 L'ajustement du variogramme

Il faut ensuite trouver une forme fonctionnelle adaptée au variogramme expérimental. Une première étape importante est d'obtenir une représentation lissée du variogramme, qui peut être obtenue avec la fonction `variog` de *geoR*. Comme toujours, ce lissage est tributaire du choix de la fenêtre. La représentation lissée est considérée parfois comme suffisante pour estimer "à l'oeil" le

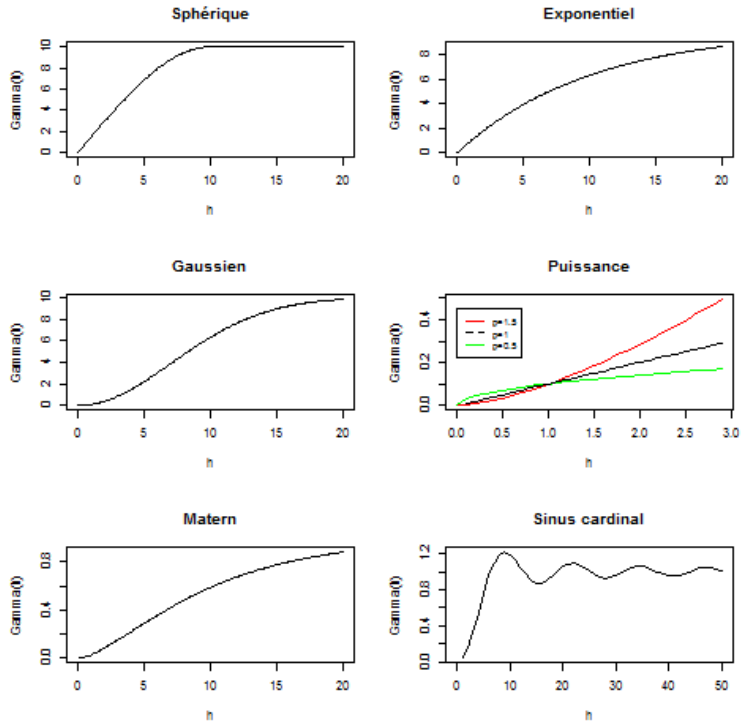


FIGURE 5.10 – Quatre exemples de variogrammes théoriques

variogramme. Cette approche est critiquée comme trop empirique par de nombreux géostatisticiens mais cette étape peut donner quelques indications, notamment sur le comportement à l'origine.

On trouvera dans la figure 5.11 trois exemples d'ajustement du variogramme expérimental des données sur les précipitations en Suisse, par un variogramme sphérique, un variogramme exponentiel sans pépite et un variogramme exponentiel avec pépite. Ces ajustements sont effectués à l'aide de la fonction `lines.variomodel` de *geoR* (RIBEIRO JR et al. 2006). C'est une première approche "à l'oeil". On voit que si le variogramme exponentiel avec pépite s'ajuste apparemment mieux aux données que le variogramme sans pépite, l'introduction de cet effet à très courte distance n'a pas de justification physique dans le cas des précipitations.

```
library(geoR)
vario.ex<-  variog(sic.100,option="bin")
vario.sphe<-(variofit(vario.ex,cov.model=  "spher",
ini.cov.pars=c(15000,200)))
par(mfrow=c(2,2), mar=c(3,3,1,1), mgp =c (2,1,0))
plot(vario.ex,main="Sphérique")
lines.variomodel(cov.model="sphe",cov.pars=c(15000,100),
nug=0,max.dist=350)
plot(vario.ex,main="Exponentiel")
lines.variomodel(cov.model="exp",cov.pars=c(15000,100),
nug=0,max.dist=350)
plot(vario.ex,main="Exponentiel avec pépite")
lines.variomodel(cov.model="exp",cov.pars=c(10000,100),
nug=5000,max.dist=350)
```

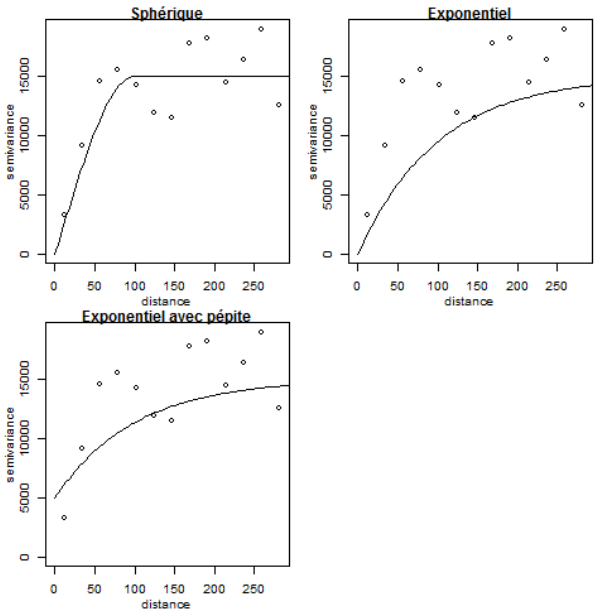


FIGURE 5.11 – Trois exemples d'ajustement du variogramme expérimental
Source : *Swiss rainfall du package geoR*

```
plot(vario.ex,main="Exponentiel avec pépite")
lines.variomodel(cov.model="matern",cov.pars=c(10000,100),
nug=0,max.dist=350,kappa=0.5)
```

Le choix est donc un compromis comme le rappellent WALLER et al. 2004 : "Même si un modèle particulier est jugé meilleur pour un ensemble de données particulier par une méthode statistique d'ajustement, ce n'est peut-être pas forcément le meilleur choix. Par exemple, le modèle gaussien est souvent sélectionné par un critère d'ajustement automatique, mais cela fournit un lissage qui apparaît souvent irréaliste. En fin de compte, le choix final du modèle devrait refléter à la fois le résultat de la procédure d'ajustement au modèle statistique et une interprétation cohérente avec la compréhension scientifique du processus à l'étude".

De nombreuses méthodes sont proposées pour ajuster le variogramme. On trouve des méthodes fondées sur les moindres carrés, ordinaires ou pondérés, des méthodes fondées sur la vraisemblance, ainsi que des méthodes bayésiennes. Dans *geoR*, les fonctions utilisées sont `variofit` (moindres carrés) et `likfit` (maximum de vraisemblance). Les méthodes sont assez techniques et nécessiteraient des développements importants. On peut renvoyer à RIBEIRO JR et al. 2006 pour une illustration de ces méthodes à partir de données simulées. Le code R est fourni dans l'article.

Ajustement par les moindres carrés ordinaires (MCO)

On cherche le vecteur des paramètres de la fonction qui minimise une fonction d'objectif simple, la somme des carrés des distances entre la valeur du semi-variogramme expérimental et celle du variogramme théorique.

$$\hat{\theta}_{MCO} = \underset{\theta \in \Theta}{argmin} \sum_{i=1}^{k} (\hat{\gamma}(h_i) - \gamma(h_i;\theta))^2 \tag{5.21}$$

Ajustement par les moindres carrés pondérés (MCP)

Les moindres carrés ordinaires ne tiennent pas compte du nombre de couples de points qui interviennent dans le calcul de chacun des points du variogramme expérimental, contrairement aux moindres carrés pondérés.

$$\hat{\theta}_{MCP} = \underset{\theta \in \Theta}{argmin} \sum_{i=1}^{k} \frac{\#N(h_i)}{\gamma(h_i;\theta)^2} \left(\hat{\gamma}(h_i) - \gamma(h_i;\theta)\right)^2 \tag{5.22}$$

Ajustement par les moindres carrés généralisés (MCG)

D'autres auteurs ont proposé les moindres carrés généralisés pour prendre en compte l'hétéroscédasticité.

$$\hat{\theta}_{MCG} = \underset{\theta \in \Theta}{argmin} \left(\hat{\gamma}_n - \gamma(\theta)\right)^T Cov(\gamma_n)^{-1} \left(\hat{\gamma}_n - \gamma(\theta)\right) \tag{5.23}$$

où γ est le vecteur $(\gamma_1, \gamma_2, ..., \gamma_K)$.

Ajustement par le maximum de vraisemblance (MV)

Les paramètres du modèle sont estimés en calculant la vraisemblance. Dans le cas non gaussien, les estimations manquent de robustesse. Les calculs sont lourds, et il faut réserver cette méthode à de petits échantillons. De plus, cette méthode requiert la stationnarité au second ordre et ne peut s'appliquer à des variogrammes non bornés. Dans ce dernier cas, il faut utiliser les moindres carrés pondérés.

5.4 Le krigeage ordinaire

5.4.1 Principe

Le terme de krigeage est dû à Georges Matheron, et fait référence aux travaux pionniers de Danie Krige, ingénieur sud-africain. Le krigeage est une méthode d'interpolation très puissante. Les exemples fournis ici sont très élémentaires. Les applications à la recherche minière ou géologique fournissant de nombreux exemples où l'on s'intéresse à l'estimation de volumes et pas seulement à des interpolations simples.

On ne présentera pas ici le krigeage *simple*, qui suppose connue la valeur de la moyenne, mais le krigeage qualifié d'*ordinaire*, qui constitue le point d'orgue de la géostatistique. Dans le krigeage ordinaire, la valeur moyenne n'est pas connue. On peut en trouver des utilisations simples dans l'interpolation des températures (JOLY et al. 2009) ou dans des études sur la qualité de l'air (LLOYD et al. 2004).

Supposons que $Z(.)$ est intrinsèquement stationnaire, que son variogramme $\gamma(h)$ est connu mais que sa moyenne m est inconnue. On dispose d'un ensemble de données $Z = [Z(s_1), .., Z(s_i), .., Z(s_N)]^t$. On veut prédire la valeur de $Z(.)$ en un point inobservé et calculer $Z(s_0)$. L'estimateur du krigeage ordinaire va être défini comme une combinaison linéaire des observations.

$$Z_{OK}(s_0) = \sum_{i=1}^{N} \lambda_i Z(s_i) \tag{5.24}$$

Les valeurs des poids λ_i ne sont pas calculées à partir d'une fonction de la distance mais en utilisant le semi-variogramme et deux critères statistiques : l'absence de biais et la minimisation de

l'erreur quadratique moyenne de prédiction. L'absence de biais implique que l'on ait l'équation 5.25 :

$$\mathbb{E}\left[\hat{Z}_{OK}(s_0)\right] = \mathbb{E}\left[Z(s_0)\right] = m \tag{5.25}$$
$$\mathbb{E}\left[\hat{Z}_{OK}(s_0)\right] = \sum_{i=1}^{N} \lambda_i \mathbb{E}[Z(s_i)] = \sum_{i=1}^{N} \lambda_i m \quad \rightarrow \quad \sum_{i=1}^{N} \lambda_i = 1$$

On va donc, en utilisant la méthode des multiplicateurs de Lagrange, minimiser $\mathbb{E}\left[\hat{Z}_{OK}(s_0) - Z(s_0)\right]^2$ sous la contrainte $\sum_{i=1}^{N} \lambda_i = 1$.

On trouvera dans l'encadré 5.4.1 la façon d'introduire le variogramme et d'aboutir aux équations de krigeage.

$$\begin{aligned} &\sum_{j=1}^{N} \lambda_j \gamma(s_i - s_j) + m = \gamma(s_0 - s_i) \\ &\sum_{i=1}^{N} \lambda_i = 1 \end{aligned} \tag{5.26}$$

La valeur de $\hat{Z}_{OK}(s_0)$ est déterminée par des points qui dépendent de la corrélation entre le point d'estimation et les points d'observation, mais aussi des corrélations entre les points d'observation. On convient en général d'écrire ces équations de krigeage sous forme matricielle :

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ . \\ . \\ \lambda_N \\ m \end{bmatrix} = \begin{bmatrix} \gamma(s_1 - s_1) & .. & \gamma(s_1 - s_N) & 1 \\ \gamma(s_2 - s_1) & .. & \gamma(s_2 - s_N) & 1 \\ . & .. & . & . \\ . & .. & . & . \\ \gamma(s_N - s_1) & .. & \gamma(s_N - s_N) & 1 \\ 1 & .. & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma(s_0 - s_1) \\ \gamma(s_0 - s_2) \\ . \\ . \\ \gamma(s_0 - s_N) \\ 1 \end{bmatrix} \tag{5.27}$$

ou de façon plus ramassée :

$$\lambda_0 = \Gamma_{ij}^{-1} \gamma_0. \tag{5.28}$$

La matrice Γ ne dépend pas du point d'estimation, et n'a donc pas à être recalculée à chaque fois. Les valeurs de tous les $\gamma(s_i - s_j)$ et $\gamma(s_0 - s_i)$ sont calculées à partir des valeurs du variogramme estimé. L'erreur quadratique moyenne de prédiction, connue sous le nom de *variance de krigeage*, est égale à $\lambda_0^t \gamma_0$. En résumé, le krigeage fournit un estimateur sans biais, de variance minimale, qui est aussi un interpolateur exact puisqu'il redonne pour chaque point connu une valeur estimée égale à la valeur observée.

Encadré 5.4.1 — Estimation des équations de krigeage. En utilisant la méthode des multiplicateurs de Lagrange, on minimise :
$\mathbb{E}\left[\hat{Z}_{OK}(s_0) - Z(s_0)\right]^2$ sous la contrainte $\sum_{i=1}^{N} \lambda_i = 1$.

On va chercher les $\lambda_1, ..., \lambda_N$ et le multiplicateur m qui permettent d'introduire la contrainte. La fonction d'objectif s'écrit donc :

$$\mathbb{E}\left[\left(\sum_{i=1}^{N} \lambda_i Z(s_i) - Z(s_0)\right)^2\right] - 2m\left(\sum_{i=1}^{N} \lambda_i - 1\right). \tag{5.29}$$

Du fait de la contrainte, on peut écrire :

$$\left[\sum_{i=1}^{N}\lambda_i Z(s_i)-Z(s_0)\right]^2=-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[(Z(s_i)-Z(s_j))^2\right]+\sum_{i=1}^{N}\left[(Z(s_i)-Z(s_0))^2\right]. \quad (5.30)$$

En prenant l'espérance des expressions, on a :

$$\mathbb{E}\left[\sum_{i=1}^{N}\lambda_i Z(s_i)-Z(s_0)\right]^2=-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j\mathbb{E}\left[(Z(s_i)-Z(s_j))^2\right]+\sum_{i=1}^{N}\lambda_i\mathbb{E}\left[(Z(s_i)-Z(s_0))^2\right]. \quad (5.31)$$

Cette expression permet de faire apparaître le variogramme et on peut réécrire la contrainte comme :

$$-\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j\gamma(s_i-s_j)+2\sum_{i=1}^{N}\gamma(s_0-s_i)-2m\left(\sum_{i=1}^{N}\lambda_i-1\right). \quad (5.32)$$

On va minimiser cette expression en dérivant par rapport à $\lambda_1,\ldots,\lambda_N$ et m, ce qui conduit aux équations de krigeage :

$$\begin{aligned}&\sum_{j=1}^{N}\lambda_j\gamma(s_i-s_j)+m=\gamma(s_0-s_i)\\&\sum_{i=1}^{N}\lambda_i=1.\end{aligned} \quad (5.33)$$

5.4.2 Application aux données de précipitations

Données brutes

Un premier krigeage a été réalisé à partir des données brutes en utilisant un modèle sphérique pour le variogramme. On trouve en figure 5.12 le variogramme sphérique utilisé, après estimation des paramètres par le maximum de vraisemblance comparé au variogramme expérimental.

```
library(geoR)
vario.ex <- variog(sic.100, bin.cloud=TRUE)
plot(vario.ex,main="")
lines.variomodel(cov.model="spher",cov.pars=c(15000,50),
nug=0,max.dist=300)
```

On peut ainsi calculer des valeurs krigées sur un carroyage, ainsi que les variances de krigeages qui sont représentées dans la figure 5.13. Les valeurs sont représentées selon les conventions de sémiologie graphique, les couleurs chaudes correspondant aux valeurs élevées.

```
library(geoR)
pred.grid <-  expand.grid(seq(0,350, l=51), seq(0,220, l=51))
rgb.palette <- colorRampPalette(c("blue", "lightblue",
"orange", "red"),space = "rgb")
kc <- krige.conv(sic.100, loc = pred.grid,
 krige=krige.control(cov.model="spherical",cov.pars=c(15000,50)))
```

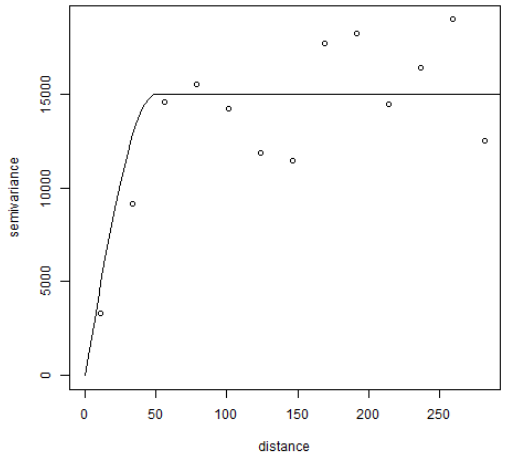


FIGURE 5.12 – Variogramme sphérique pour les données brutes
Source : *Swiss rainfall du package geoR*

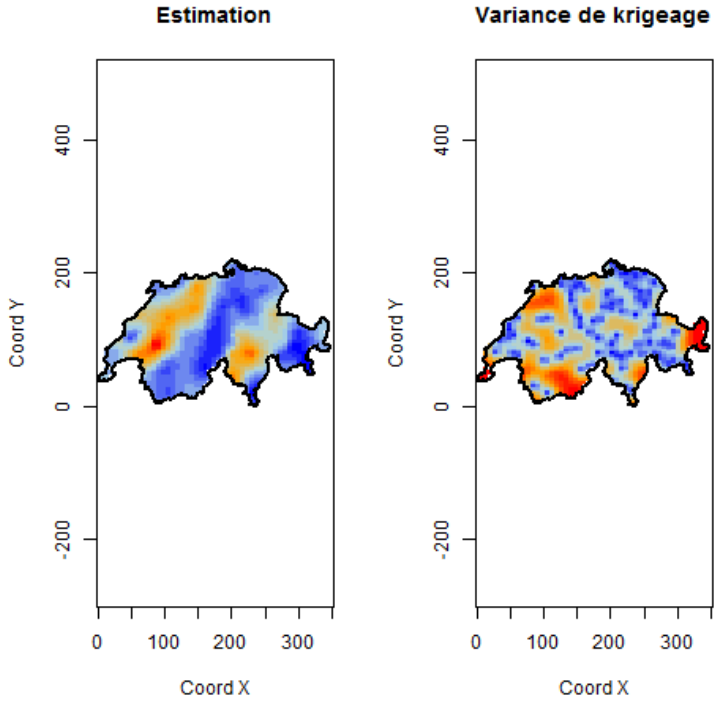


FIGURE 5.13 – Estimations et variances de krigeage pour les données brutes
Source : *Swiss rainfall du package geoR*

```
image(kc, loc = pred.grid,col =rgb.palette(20) ,xlab="Coord X",
ylab="Coord Y",borders=sic.borders,main="Estimation")
image(kc, krige.var,loc = pred.grid,col=rgb.palette(20),
xlab="Coord X",ylab="Coord Y",borders=sic.borders,
main="Variance de krigeage")
```

L'estimation a été effectuée en faisant intervenir les 100 points de l'échantillon. On peut :
— vérifier que le krigeage est sans biais sur les points observés ;
— mesurer les écarts entre valeurs estimées et valeurs observées.

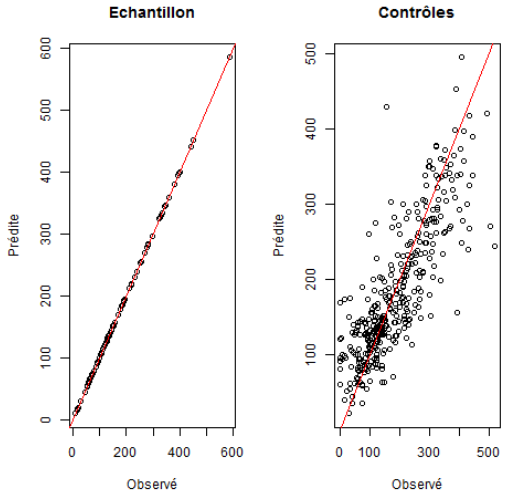


FIGURE 5.14 – Valeurs estimées et observées
Source : *Swiss rainfall du package geoR*

```
library(geoR)
kc1<-  krige.conv(sic.100, loc = sic.100$coords,
krige=krige.control(cov.model="spherical",cov.pars=c(16000,47)))
kc2<-  krige.conv(sic.100, loc = sic.367$coords,
krige=krige.control(cov.model="spherical",cov.pars=c(16000,47)))
plot(sic.100$data,kc1$predict,xlab="Observ\'e",ylab="Pr\'edite",
main="Echantillon")
abline(a=0,b=1,col="red")
plot(sic.367$data,kc2$predict,,xlab="Observ\'e",ylab="Pr\'edite",
main="Contr\^oles")
abline(a=0,b=1,col="red")
```

Données après transformations

L'histogramme de la figure 5.3 indiquait que la distribution des données de précipitations s'écartait d'une distribution gaussienne. Une première possibilité, classique en statistique, est la modification des variables. En effet, il peut être avantageux de travailler avec des données qui suivent une loi normale. Ce n'est pas absolument requis dans le modèle de krigeage, mais l'hypothèse de linéarité du krigeage n'est vraiment performante que lorsque les données sont gaussiennes. Les transformations classiques sont les transformations logarithmiques, ou de façon plus générale les

transformations de Box-Cox. La fonction `boxcoxfit` de *geoR* suggère dans le cas des données un coefficient proche de 0.5. Les données exploratoires sont représentées dans la figure 5.15.

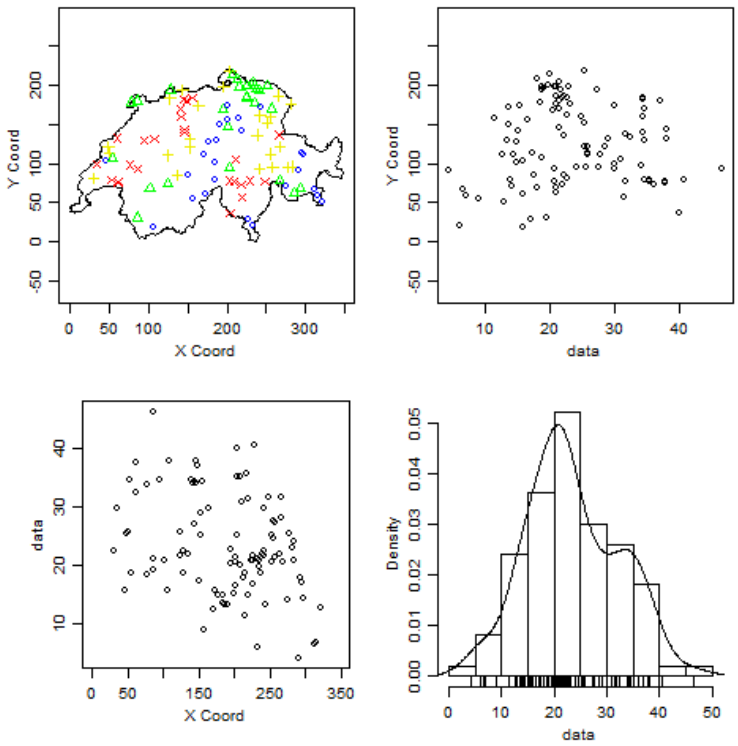


FIGURE 5.15 – Données exploratoires, après transformation Box-Cox
Source : *Swiss rainfall du package geoR*

```
library(geoR)
plot.geodata(sic.100,bor=sic.borders,lambda=0.5)
```

Les données sur les précipitations ont souvent été étudiées. RIBEIRO JR et al. 2004 préconisent la même transformation des variables. Pour le variogramme, ils proposent l'utilisation d'un modèle de Matern pour lequel $K = 1$. Les paramètres du modèle sont déterminés à l'aide du maximum de vraisemblance. Les variogrammes expérimentaux et théoriques pour les données transformées sont présentés en figure 5.16.

```
library(geoR)
vario.ext<-  variog(sic.100,option="bin",lambda=0.5)
plot(vario.ext)
lines.variomodel(cov.m = "mat",cov .p =c (105, 36), nug = 6.9,
                 max.dist = 300,kappa  = 1, lty = 1)
```

Comme pour les données brutes, on peut fournir une cartographie des estimations et des valeurs de krigeage (figure 5.17).

```
library(geoR)
kct<-  krige.conv(sic.100, loc = pred.grid,
krige=krige.control(cov.model="matern",cov.pars=c(105, 36),
kappa=1,nugget=6.9,lambda=0.5))
```

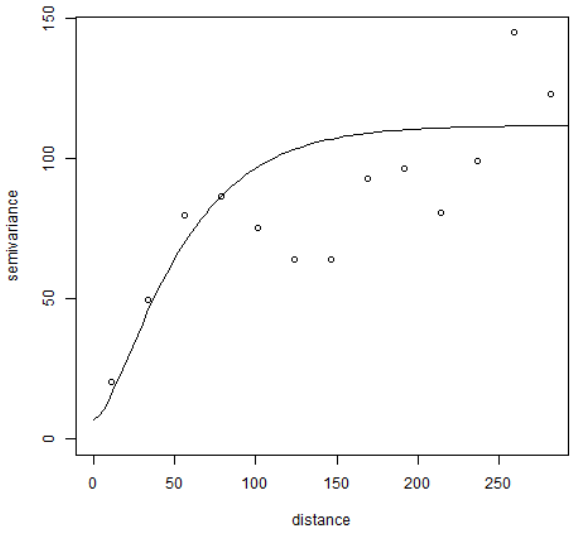


FIGURE 5.16 – Variogramme expérimental et variogramme théorique après transformation
Source : *Swiss rainfall du package geoR*

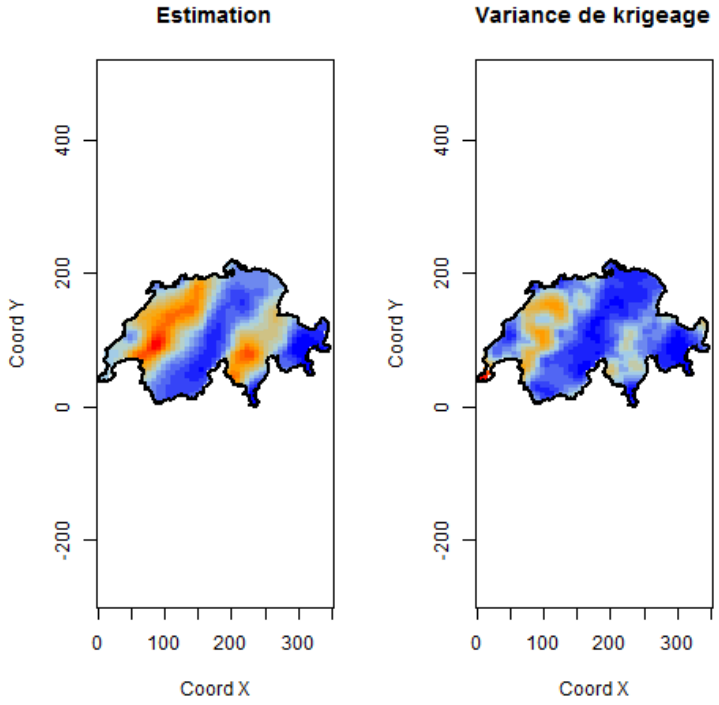


FIGURE 5.17 – Estimations et variance de krigeage des précipitations en Suisse après transformation
Source : *Swiss rainfall du package geoR*

```
pred.grid <-  expand.grid(seq(0,350, l=51),seq (0,220, l=51))
rgb.palette <- colorRampPalette(c("blue", "lightblue",
"orange", "red"),space = "rgb")
image(kct, loc = pred.grid,col =rgb.palette(20) , xlab="Coord X",
ylab="Coord Y",borders=sic.borders,main="Estimation")
image(kct, krige.var,loc = pred.grid,col =rgb.palette(20) ,
xlab="Coord X",ylab="Coord Y",borders=sic.borders,
main="Variance de krigeage")
```

On peut comparer les valeurs estimées et les valeurs observées (5.18).

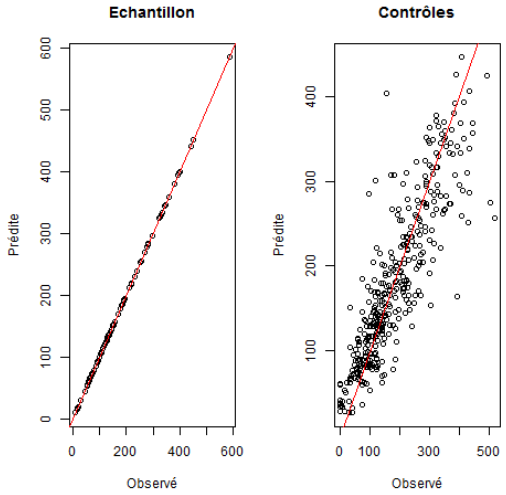


FIGURE 5.18 – Valeurs estimées et observées
Source : *Swiss rainfall du package geoR*

```
library(geoR)
kct1<-  krige.conv(sic.100, loc = sic.100$coords,
krige=krige.control(cov.model="spherical",cov.pars=c(16000,47),
kappa=1,nugget=6.9,lambda=0.5))
kct2<-  krige.conv(sic.100, loc = sic.367$coords,
krige=krige.control(cov.model="spherical",cov.pars=c(16000,47),
kappa=1,nugget=6.9,lambda=0.5))
plot(sic.100$data,kct1$predict,xlab="Observ\'e",ylab="Pr\'edite",
main="Echantillon")
abline(a=0,b=1,col="red")
plot(sic.367$data,kct2$predict,,xlab="Observ\'e",ylab="Pr\'edite",
main="Contr\^oles")
abline(a=0,b=1,col="red")
```

Si on prend comme critère la racine carrée de l'écart quadratique moyen

$$RMSE(y) = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2}{n}},$$

on constate une amélioration de la prédiction si on utilise les données transformées modélisées par

le modèle de Matern. Dans le cas des données brutes la RMSE vaut 62.3, dans le cas des données transformées 55.2.

5.5 Support et changement de support

La question des échelles d'analyse est de première importance en analyse spatiale. Les géographes, depuis Openshaw notamment utilisent le terme de *Modifiable areal unit problem* (MAUP), présenté dans le chapitre 1 : "Analyse spatiale descriptive". On peut résumer le MAUP comme la superposition d'un effet de zonage et d'un effet d'agrégation. On peut l'illustrer de façon simple à partir de données surfaciques, en montrant la variabilité des résultats obtenus selon le découpage territorial utilisé.

En géostatistique, on parle de *change of support problem* (COSP). Ce problème est de fait plus général que le MAUP, puisqu'il renvoie à la taille, à la forme et aux orientations. Le COSP en géostatistique prend son origine dans des préoccupations très pratiques, liées à la recherche minière dont on peut trouver les origines dans les travaux de Krige, que développera Matheron. Les articles de Krige sont contemporains de ceux de Yule et Kendall en statistique classique, qui anticipent ceux d'Openshaw en géographie. Dans une perspective pratique, les géostatisticiens avaient rapidement perçu qu'il était plus important de prédire une valeur sur un bloc important que sur un point, quand bien même la première prédiction dérivait de la seconde.

Le support peut être un point, un bloc plus ou moins gros, une réunion de points dans une configuration géométrique donnée. Il y a des relations entre les valeurs prises sur des volumes différents. Dans le cadre de variables additives, pour un volume V partitionné en unités v_i de même support v (V étant un multiple de v) on a :

$$z(V) = \frac{1}{n}\sum_{i=1}^{n} z(v_i) \tag{5.34}$$

ou si v est ponctuel :

$$z(V) = \frac{1}{V}\int_V z(x)dx. \tag{5.35}$$

$Z(V)$ est qualifiée de variable *régularisée*, parce qu'elle accroît la régularité statistique. En fait, c'est une forme particulière de variable régularisée, la forme plus générale étant une *convoluée de* Z (CHILES et al. 2009).

5.5.1 Variance de dispersion empirique et relations d'additivité de Krige

Définition 5.5.1 — Variance de dispersion empirique.

$$s^2(v|V) = \frac{1}{n}\sum_{i=1}^{n} [z(v_i) - z(V)]^2 \tag{5.36}$$

Si V est inclus dans un domaine plus large noté D, on peut démontrer la relation 5.37, appelée **relation d'additivité de Krige** :

$$s^2(v|D) = s^2(v|V) + s^2(V|D). \tag{5.37}$$

ARMSTRONG 1998 donne un exemple pédagogique à partir des rendements en millet sur 16 blocs de 2m de côté, partagés en 64 parcelles de 1m de côté. La moyenne vaut 201 dans le cas des blocs comme dans le cas des parcelles. La variance de dispersion des blocs dans le champ vaut 16.64, celle des parcelles 27.59, ce qui veut dire que la dispersion des parcelles dans les blocs vaut 10.85 (voir table 5.1).

735	325	45	140	125	175	167	485
540	420	260	128	20	30	105	70
450	200	337	190	95	260	245	278
180	250	380	405	250	80	515	605
124	120	430	175	230	120	460	260
40	135	240	35	190	135	160	170
75	95	20	35	32	95	20	450
200	35	100	59	2	45	58	90

505	143	88	207
270	328	171	411
102	220	154	263
101	54	44	155

TABLE 5.1 – Valeurs observées sur les parcelles (en haut) et les blocs (en bas)
Source : ARMSTRONG *1998*.

5.5.2 Variogramme de la variable régularisée

On peut définir la variance par bloc à partir des informations sur les données ponctuelles (fonction de covariance).

$$Var\,[Z(V)] = \overline{C}\,(V,V) = \frac{1}{|V|^2}\int_V\int_V C(x-y)\,dx\,dy \tag{5.38}$$

où C désigne la fonction de covariance pour les données ponctuelles, et $\overline{C}$ la covariance des données par bloc. La covariance s'écrit :

$$Cov\left[Z(V),Z(V')\right] = \overline{C}(V,V') = \frac{1}{|V|\,|V'|}\int_V\int_{V'} C(x-y)\,dx\,dy. \tag{5.39}$$

La fonction de covariance C_V est définie en introduisant V_h qui est le translaté du support V par le vecteur h :

$$C_V(h) = Cov\,[Z(V),Z(V_h)] = \overline{C}(V,V_h). \tag{5.40}$$

On peut également déduire du variogramme ponctuel le variogramme de la variable régularisée :

$$\gamma_V(h) = \overline{\gamma}(V,Vh) - \overline{\gamma}(V,V) \tag{5.41}$$

avec $\gamma(V,V) = \frac{1}{|V|^2}\int_V\int_V \gamma(x-y)\,dx\,dy$ et $\overline{\gamma}(V,V_h) = \frac{1}{|V|^2}\int_V\int_{V_h} \gamma(x-y)\,dx\,dy$ où γ est le variogramme calculé à partir des observations ponctuelles.

On peut alors montrer que $\gamma_V(h) \sim \gamma(h) - \overline{\gamma}(V,V)$ ce qui se traduit graphiquement par la figure 5.19.

Pour passer du variogramme ponctuel au variogramme régularisé, on conserve le même type de modèle théorique en corrigeant le palier et la portée.

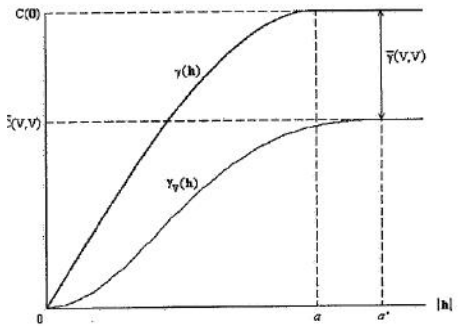


FIGURE 5.19 – Passage du variogramme ponctuel au variogramme régularisé

5.5.3 Krigeage par bloc

On peut déterminer des équations de krigeage, dans la logique du krigeage ordinaire. De $\mathbb{E}[Z(x)] = m$, on en déduit que $\mathbb{E}[Z_V] = m$.

On cherche comme dans le cas du krigeage ordinaire un estimateur qui soit une combinaison linéaire des observations collectées.

$$Z_V = \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{5.42}$$

La démarche est la même que celle qui est exposée en encadré 5.4.1 (minimisation sous contrainte) et on aboutit aux équations de krigeage :

$$\begin{aligned} &\sum_{j=1}^{n} \lambda_j \gamma(x_i, x_j) + \mu = \gamma(x_i, V) \quad pour \quad i = 1, 2, ..., n \\ &\sum_{i=1}^{n} \lambda_i = 1 \end{aligned} \tag{5.43}$$

avec $\overline{\gamma}(x_i, V) = \frac{1}{|V|} \int_V \gamma(x_i - x)\, dx$.

5.6 Extensions

Le krigeage ordinaire constitue la méthode de base de la géostatistique. Mais elle n'en constitue au final qu'une partie, et de nombreux développements ont vu le jour, notamment à l'école des Mines de Fontainebleau.

Les hypothèses de stationnarité intrinsèque restent assez restrictives, notamment la constance de la moyenne. De nombreuses méthodes ont été élaborées pour permettre d'introduire des hypothèses moins contraignantes, ou pour utiliser des informations auxiliaires. Dans ce paragraphe, on présentera quelques modifications et variantes, à partir essentiellement des données suisses sur les précipitations, et d'un autre jeu de données fréquemment mis à contribution, sur les teneurs en divers minéraux dans une boucle de la Meuse.

5.6.1 Le cokrigeage

La géostatistique a développé depuis longtemps des méthodes multivariées. L'une d'entre elle est le cokrigeage, qui va prendre en compte plusieurs variables. La définition qu'en donne WALLER et al. 2004 est la suivante :

Définition 5.6.1 — cokrigeage. "Le cokrigeage est une extension du krigeage au cas de deux variables spatiales ou plus. Il a été développé à l'origine comme une technique pour améliorer la prédiction d'une variable pour laquelle seuls quelques échantillons pourraient être prélevés, en utilisant sa corrélation spatiale avec d'autres variables plus facilement mesurées. Le cokrigeage diffère du krigeage avec dérive externe en ce sens que les variables explicatives ne sont pas supposées indiquer la nature d'une tendance dans la variable primaire, mais sont elles-mêmes des variables aléatoires spatiales pour lesquelles on peut utiliser l'analyse variographique".

On peut définir un covariogramme croisé :

$$\gamma_{ZY}(h) = \frac{1}{2p(h)} \sum_{i=1}^{p(h)} (z(s_i) - z(s_i+h))\,(y(s_i) - y(s_i+h)) \tag{5.44}$$

avec $p(h) = Card\left\{(s_i, s_j) \middle| \left|s_i - s_j\right| \approx h\right\}$

Comme pour le krigeage, il y aura plusieurs versions pour le cokrigeage. On ne présentera que le cokrigeage ordinaire. On se limitera au cas où l'on n'introduit qu'une variable auxiliaire, que l'on notera Y. L'estimateur que l'on calcule est de la forme :

$$Z(s_0) = \sum_{i=1}^{n_Z} \lambda_i Z(s_i) + \sum_{i=1}^{n_Y} \alpha_i Y(s_i) \tag{5.45}$$

avec les contraintes d'absence de biais :

$$\begin{aligned} &\sum_{i=1}^{n_Z} \lambda_i = 1 \\ &\sum_{i=1}^{n_Y} \alpha_i = 0. \end{aligned} \tag{5.46}$$

Sous forme matricielle, les équations de cokrigeage s'écrivent :

$$\begin{bmatrix} \Gamma_{ZZ} & \Gamma_{ZY} & 1 & 0 \\ \Gamma_{YZ} & \Gamma_{YY} & 0 & 1 \\ 1' & 0' & 0 & 0 \\ 0 & 1' & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \alpha \\ \mu_Z \\ \mu_Y \end{bmatrix} = \begin{bmatrix} \gamma_{ZZ} \\ \gamma_{YZ} \\ 1 \\ 0 \end{bmatrix} \tag{5.47}$$

On illustrera cette méthode à l'aide des données *Meuse* fournies dans le package *sp* et étudiées notamment par Pebesma (PEBESMA 2001) et Rossiter (ROSSITER 2017). Ce dernier fournit sur son site des programmes R.

■ **Exemple 5.1** Analyse par cokrigeage des données *Meuse*

Les données *Meuse* fournissent des mesures localisées de teneur en plomb, zinc, cadmium, mais aussi d'autres variables comme l'altitude, la teneur en matière organique du sol. Le package *sp* contient une table nommée *Meuse*, pouvant être chargée avec la fonction : `data(meuse)`. Il fournit également une grille de 40 × 40 m : *meuse.grid* et les limites du département : *meuse.riv*.

Les données sont constituées de 155 observations sur des supports de 15x15 m, sur les 20 cm supérieurs des sols alluviaux de la rive droite de la Meuse. Les données associées fournissent les coordonnées géographiques des observations, leur altitude, et les concentrations en cadmium, cuivre, plomb, zinc, matière organique. On trouve aussi la distance à la Meuse et la fréquence

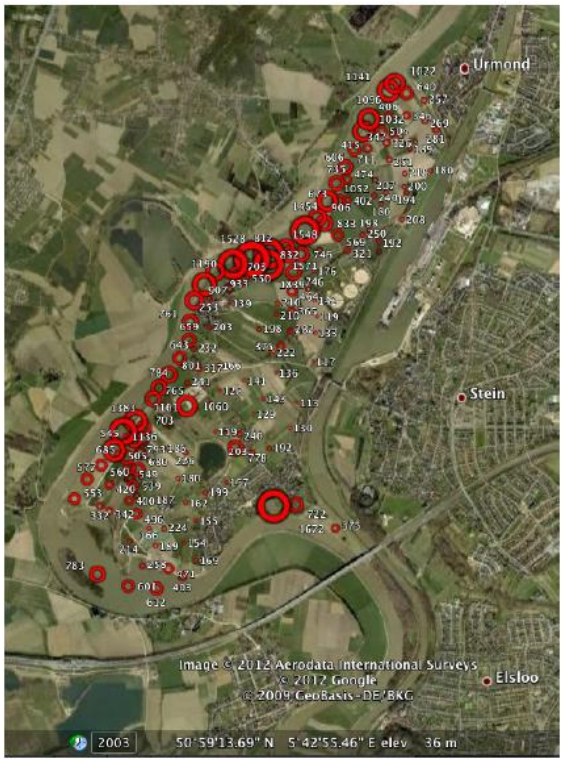


FIGURE 5.20 – Géographie et position des prélèvements
Source : *Données Meuse du package sp*

des inondations. Dans l'exemple fourni par ROSSITER 2017, c'est la teneur en plomb qui est étudiée (après une transformation logarithmique) et la teneur en matière organique qui va servir de covariable.

L'analyse variographique dans le cas du cokrigeage repose sur l'étude des deux variogrammes simples et du variogramme croisé.

Le même exercice est effectué en prenant comme covariable le logarithme de la teneur en zinc. La figure ci-dessous présente les résultats de krigeage de la teneur en plomb, par krigeage ordinaire et par cokrigeage en utilisant les deux variables mentionnées ci-dessus (teneur en zinc et matière organique). On trouve dans la figure 5.22 les valeurs krigées (colonne de gauche) et les valeurs résiduelles (colonne de droite). On trouve successivement le krigeage ordinaire, le cokrigeage avec la matière organique comme covariable, puis le cokrigeage avec la teneur en zinc.

Le code nécessaire pour ce traitement est assez long. Le lien ci-dessous permet d'accéder au programme R mis à disposition par Rossiter (ROSSITER 2007) :

`http://www.css.cornell.edu/faculty/dgr2/teach/R/ck_plotfns.R`.

Si l'on compare, en utilisant la RMSE, les trois modélisations, on trouve les résultats suivants :

— 0.166 pour le krigeage ordinaire ;
— 0.226 pour le cokrigeage avec la teneur en matière organique ;
— 0.078 pour le cokrigeage avec la teneur en zinc.

Le cokrigeage faisant intervenir la teneur en zinc améliore donc les performances du krigeage ordinaire. Ce n'est pas le cas pour le cokrigeage avec la teneur en matière organique. ∎

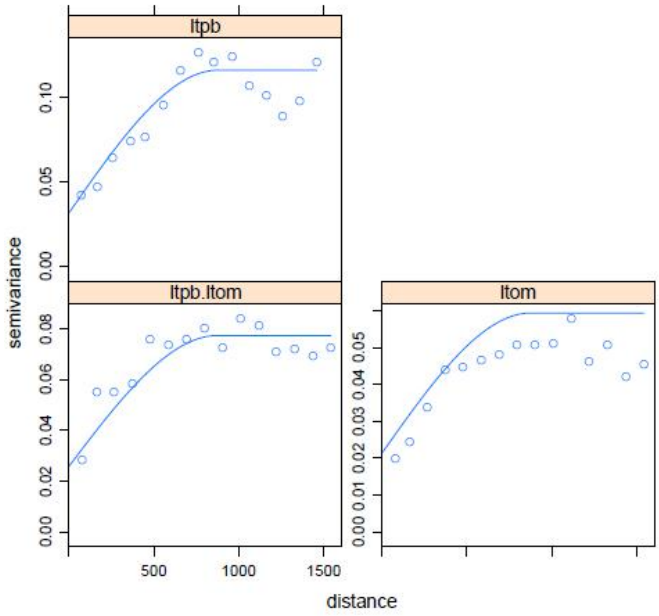


FIGURE 5.21 – Variogrammes simples et croisés
Source : *Données Meuse du package sp*

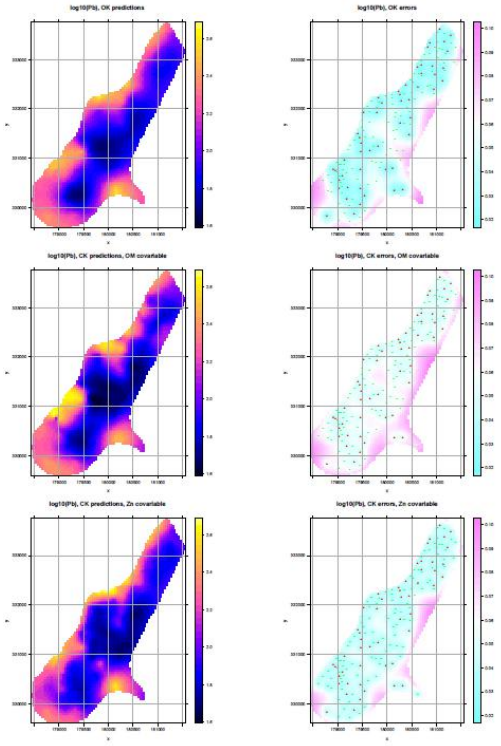


FIGURE 5.22 – Estimation et variance de krigeage et cokrigeage
Source : *Données Meuse du package sp*

5.6.2 Le krigeage universel

Dans de nombreux cas, la valeur moyenne n'est pas constante, et on ne peut pas utiliser le krigeage ordinaire. C'est le cas lorsque l'on observe des relations déterministes entre la valeur de la variable et sa position dans l'espace. La variable régionalisée peut alors s'écrire :

$$Z(s) = m(s) + Y(s) \tag{5.48}$$

où $m(s)$ représente la composante déterministe.

■ **Exemple 5.2 — Analyse par cokrigeage des données *Meuse*.** Les données *Meuse* fournissent deux variables susceptibles de construire une composante déterministe : la distance à la rivière et la fréquence des inondations (figure 5.23). Elles sont de nature différente des covariables utilisées précédemment dans le cokrigeage.

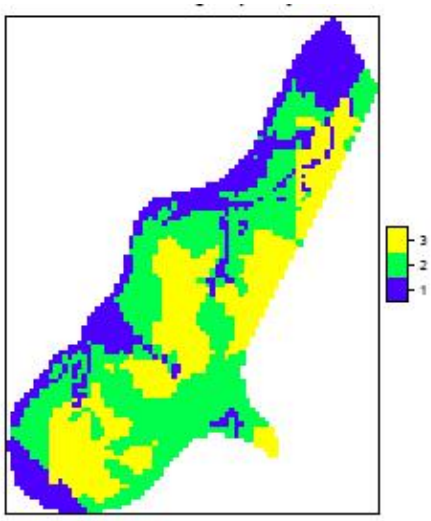


FIGURE 5.23 – Fréquence des inondations
Source : *Données Meuse du package sp*

Sur son site, Rossiter fournit des exemples permettant de comparer les prédictions du krigeage ordinaire et de deux modèles de krigeage universel.

Si l'on compare, en utilisant la RMSE, les trois modélisations, on trouve les résultats suivants :
— 0.173 pour le krigeage ordinaire ;
— 0.141 pour le modèle avec fréquence des inondations ;
— 0.145 pour le modèle avec fréquence des inondations et distance à la rivière.

L'introduction de la fréquence des inondations améliore les prédictions, mais l'introduction en plus de la distance le dégrade (probablement à cause de la corrélation des deux variables). Le code R utilisé est disponible à l'adresse ci-dessous. Il constitue un bon exemple d'illustration et d'approfondissement de ce qui a été présenté dans ce chapitre.

`http://www.css.cornell.edu/faculty/dgr2/teach/R/gs_short_ex.pdf` ■

5.7 Modèles mixtes avec variogramme

On peut trouver d'autres utilités aux outils développés en géostatistique. Lorsqu'on effectue des régressions sur des données spatialisées, il est fréquent que les résidus soient spatialement autocorrélés. Cette corrélation peut être mise en évidence par l'indicateur de Moran d'autocorrélation spatiale (voir chapitre 3 : "Indices d'autocorrélation spatiale"). La prise en compte de

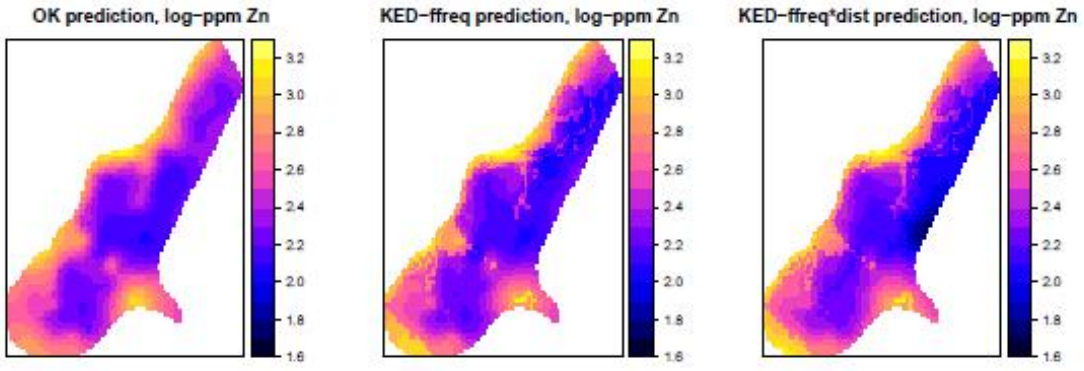


FIGURE 5.24 – Trois estimations
Source : *Données Meuse du package sp*
Note : krigeage ordinaire (à gauche), fréquence des inondations (au milieu), fréquence des inondations et distance à la rivière (à droite)

cette autocorrélation peut se faire en utilisant des modèles économétriques spatiaux (voir chapitre 6 : "économétrie spatiale : modèles courants") ou la régression géographiquement pondérée (voir chapitre 9 : "Régression géographiquement pondérée").

Lorsque les données s'y prêtent, le variogramme peut aussi être utilisé pour étudier la structure spatiale des résidus dans des modèles linéaires. Les exemples sont plus souvent présentés dans des ouvrages traitant d'écologie ou d'épidémiologie. On pourra en trouver des illustrations dans des manuels généraux de statistique spatiale comme WALLER et al. 2004 ou SCHABENBERGER et al. 2017, ainsi que des ouvrages traitant d'écologie, comme PLANT 2012 ou ZUUR et al. 2009, les deux derniers fournissant des exemples d'implémentation en R.

■ **Exemple 5.3 — Analyse de la structure spatiale des résidus avec un variogramme.** Un exemple d'utilisation sur des données écologiques est fournie par ZUUR et al. 2009 [2]. L'exemple provient de données collectées sur les forêts de la section Raifa de la biosphère naturelle d'état de Voljsko-Kamsky.

La variable d'intérêt est un indice de "boréalité" (*Bor*) défini comme la part des espèces spécifiquement boréales par rapport au nombre total des espèces sur un site. On dispose aussi de variables explicatives fournies par des images satellitaires :

1. l'indice normalisé de différence de végétation ;
2. la température ;
3. un indice d'humidité ;
4. un indice de verdure.

Du fait de la forte colinéarité entre ces variables, on n'a utilisé que l'humidité pour expliquer l'indice de boréalité. L'analyse de la variance réalisée avec les moindres carrés ordinaires fournit les résultats suivants :

Les coordonnées des sites permettent de fournir une vision exploratoire de la spatialisation des résidus du modèle MCO.

```
library(sp)
library(nlme)
```

2. https://github.com/James-Thorson/2016_class_CMR/tree/master/Other%20material/Zuur%20et%20al.%202007/ZuurDataMixedModelling

Variable	Valeur estimée	écart-type
Constante	27.63	0.981
Wet	429.609	27.45

TABLE 5.2 – Estimation par les moindres carrés

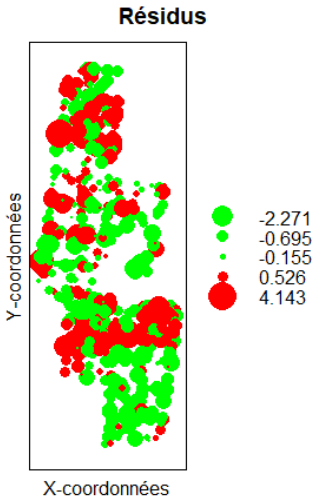


FIGURE 5.25 – Variogramme des résidus
Source : *Données Meuse du package sp*

```
Boreality<-read.table("C:/jmf/Boreality.txt",header=TRUE)
B.lm <- lm(boreal~ Wet, data  = Boreality)
E <- rstandard(B.lm)
graphique <- data.frame(E, Boreality$x, Boreality$y)
library(sp)
coordinates(graphique) <- c("Boreality.x","Boreality.y")
bubble(graphique, "E", col= c ("green","red"),main = "R\'esidus",
xlab = "X-coordonn\'ees", ylab = "Y-coordonn\'ees")
```

Le modèle MCO ne permet pas d'introduire de structure spatiale sur les résidus. On ne peut l'introduire que dans un modèle linéaire généralisé. On l'estimera à l'aide de la fonction `gls` du package *nlme* de R. Ce package contient une fonction permettant d'estimer le variogramme.
Le variogramme expérimental est représenté ci-dessous (figure 5.26)

```
mod<-gls(boreal~Wet,data=bor)
summary(mod)
plot(Variogram(mod,form=~x+y,maxdist=10000),xlim=c(0,10000))
```

L'estimation initiale du modèle, sans introduction de la structure spatiale, donne les mêmes résultats que les MCO. La commande `update` permet d'introduire une structure spatiale à l'aide d'une analyse variographique, et de réestimer le modèle.

On trouvera dans le tableau 5.3 le résultat de la comparaison entre les MCO et les modèles utilisant les variogrammes sphériques, gaussiens et exponentiels.

Les critères AIC et L montrent qu'on peut améliorer le modèle en utilisant un variogramme

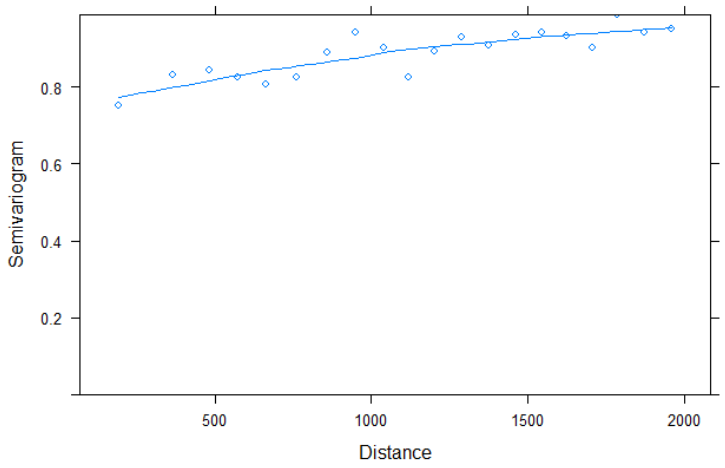


FIGURE 5.26 – Variogramme des résidus
Source : *Données Meuse du package sp*

Variogramme	AIC	Vraisemblance	L	Significativité
MCO	3855	-1924		
Sphérique	3859	-1924	1	1
Gaussien	3750	-1870	109	<0.001
Exponentiel	3740	-1865	119	<0.001

TABLE 5.3 – Critères AIC et L selon le type de variogramme

sphérique pour modéliser les résidus. Avec le variogramme exponentiel, les résultats de la régression sont présentés dans le tableau 5.4.

Variable	Valeur estimée	écart-type
constante	18.099	2.333
Wet	180.247	34.932

TABLE 5.4 – Estimations

```
f1 <- formula(boreal ~ Wet)
B1.gls <- gls(f1, data = Boreality)
Vario.gls <- Variogram(B1.gls, form =~ x + y,robust = TRUE,
maxDist = 2000,resType = "pearson")
B1A <- gls(f1, correlation = corSpher(form =~ x + y,nugget = TRUE),
data = Boreality)
B1B <- gls(f1, correlation = corLin(form =~ x + y,nugget = TRUE),
data = Boreality)
B1C <- gls(f1, correlation = corRatio(form =~ x + y,nugget = TRUE),
data = Boreality)
B1D <- gls(f1, correlation = corGaus(form =~ x + y,nugget = TRUE),
data = Boreality)
B1E <- gls(f1, correlation = corExp(form =~ x + y,nugget = TRUE),
data = Boreality)
AIC(B1.gls, B1A, B1B, B1C, B1D, B1E)
B1 <- lm(f1, data = Boreality)
anova(B1.gls,B1A)
anova(B1.gls,B1D)
anova(B1.gls,B1E)
summary(B1E)
```

Les paramètres du modèle sont significatifs. L'influence de l'humidité est moins marquée lorsqu'on introduit l'autocorrélation spatiale dans le modèle. ∎

Conclusion

Le premier chapitre de ce manuel présente les trois grands domaines de la statistique spatiale adaptés à l'analyse des données continues, surfaciques ou ponctuelles. La géostatistique, utilisée pour les données continues, est moins directement liée aux travaux de la statistique publique. Il semblait néanmoins utile d'en faire une présentation rapide dans le manuel. D'un point de vue pédagogique, les méthodes géostatistiques illustrent particulièrement bien comment la prise en compte de l'autocorrélation spatiale (à travers le variogramme) permet d'améliorer les estimateurs. D'un point de vue plus opérationnel, sans rentrer dans la complexité des travaux de recherche minière, la géostatistique, *via* les méthodes de krigeage, est utile pour modéliser des données continues plus simples (données climatiques par exemple). L'école des Mines de Fontainebleau, qui a joué un rôle décisif dans le développement de ces méthodes, a utilisé un langage un peu inhabituel pour les statisticiens, mais des échanges nombreux ont eu lieu, depuis les travaux de Cressie, pour faire le lien entre les approches des uns et des autres. Le livre classique de Chilès et Delfiner en est un bon exemple (CHILES et al. 2009). Dans le domaine de la santé, des travaux importants ont mobilisé

les méthodes géostatistiques pour la modélisation de données épidémiologiques, notamment ceux de Diggle (DIGGLE et al. 2003) davantage connu pour ses travaux sur les méthodes ponctuelles. Pour finir, on ne peut que recommander au statisticien appelé un jour à utiliser des modèles la lecture de l'article de réflexion du fondateur de la géostatistique (MATHERON 1978).

Annexes

Rappels mathématiques

Les expressions des variogrammes théoriques, notamment le variogramme de Matern font appel à des expressions mathématiques peu usuelles, dont on rappelle l'expression ci-dessous (CHILES et al. 2009).

La fonction Gamma

$\Gamma(x) = \int_0^\infty e^{-u} u^{x-1} du$

Dans le cas de valeurs entières :

$\Gamma(n+1) = n!$

Les fonctions de Bessel

La fonction de Bessel du premier type est la suivante :

$J_\nu(x) = \left(\frac{x}{2}\right)^\nu \sum_{k=0}^\infty \frac{(-1)^k}{k!\Gamma(\nu+k+1)} \left(\frac{x}{2}\right)^{2k}$

La fonction de Bessel modifiée du premier type est la suivante :

$I_\nu(x) = \left(\frac{x}{2}\right)^\nu \sum_{k=0}^\infty \frac{1}{k!\Gamma(\nu+k+1)} \left(\frac{x}{2}\right)^{2k}.$

La fonction de Bessel modifiée du deuxième type est définie à partir de la précédente :

$K_\nu(x) = \frac{\pi}{2} \frac{I_{-\nu}(x) + I_\nu(x)}{sin\pi\nu}.$

Références - Chapitre 5

ARMSTRONG, Margaret (1998). *Basic linear geostatistics*. Springer Science & Business Media.

CHAUVET, Pierre (2008). *Aide-mémoire de géostatistique linéaire*. Presses des MINES.

CHILES, Jean-Paul et Pierre DELFINER (2009). *Geostatistics : modeling spatial uncertainty*. T. 497. John Wiley & Sons.

DIGGLE, Peter J, Paulo J RIBEIRO JR et Ole F CHRISTENSEN (2003). « An introduction to model-based geostatistics ». *Spatial statistics and computational methods*. Springer, p. 43–86.

JOLY, Daniel et al. (2009). « Interpolation par régressions locales : application aux précipitations en France ». *L'Espace géographique* 38.2, p. 157–170.

LLOYD, Christopher D et Peter M ATKINSON (2004). « Increased accuracy of geostatistical prediction of nitrogen dioxide in the United Kingdom with secondary data ». *International Journal of Applied Earth Observation and Geoinformation* 5.4, p. 293–305.

MATHERON, Georges et al. (1965). *Les variables régionalisées et leur estimation*. Masson et Cie.

MATHERON, Georges (1978). *Estimer et choisir : essai sur la pratique des probabilités*. Ecole nationale supérieure des mines de Paris.

PEBESMA, Edzer J (2001). « Gstat user's manual ». *Dept. of Physical Geography, Utrecht University, Utrecht, The Netherlands*.

PLANT, Richard E (2012). *Spatial data analysis in ecology and agriculture using R*. cRc Press.

RIBEIRO JR, Paulo J et Peter J DIGGLE (2006). « geoR : Package for Geostatistical Data Analysis An illustrative session ». *Artificial Intelligence* 1, p. 1–24.

RIBEIRO JR, Paulo Justiniano et Peter J DIGGLE (2004). « Model Based Geostatistics ». *Springer Series in Statistics*.

ROSSITER, David G (2017). « An introduction to geostatistics with R/gstat Version 3.7, 12-May-2017. »

ROSSITER, DG (2007). « Co-kriging with the gstat package of the R environment for statistical computing ». *Web : http ://www. itc. nl/rossiter/teach/R/R ck. pdf*.

SCHABENBERGER, Oliver et Carol A GOTWAY (2017). *Statistical methods for spatial data analysis*. CRC press.

WALLER, Lance A et Carol A GOTWAY (2004). *Applied spatial statistics for public health data*. T. 368. John Wiley & Sons.

ZUUR, AF et al. (2009). « Mixed effects models and extensions in ecology with R. Gail M, Krickeberg K, Samet JM, Tsiatis A, Wong W, editors ». *New York, NY : Spring Science and Business Media*.

Partie 3 : Prendre en compte les effets spatiaux

6. Économétrie spatiale : modèles courants

JEAN-MICHEL FLOCH
Insee

RONAN LE SAOUT
Ensai

Résumé

Ce chapitre décrit la conduite d'une étude d'économétrie spatiale, en s'appuyant sur une modélisation descriptive du taux de chômage par zone d'emploi. Les modèles spatiaux ont néanmoins une application plus large, l'approche étant compatible avec tout problème où des relations de "voisinage" interviennent. La théorie économique caractérise en effet de nombreux cas d'interactions entre agents (produits, entreprises, individus), qui ne sont pas nécessairement de nature géographique. Le chapitre se concentre sur l'étude de la corrélation spatiale, et donc sur ces différentes interactions, et aborde les liens avec l'hétérogénéité spatiale, à savoir les phénomènes différenciés

spatialement. Plusieurs formes d'interactions existent, relatives à la variable à expliquer, aux variables explicatives ou aux variables inobservées. De nombreux modèles se retrouvent donc en concurrence, à partir d'une même définition préalable des relations de voisinage. Une méthodologie de choix de modèle (estimation et tests) est détaillée pas à pas. Des effets de rétroaction entraînent une interprétation particulière, et plus complexe, des résultats.

La lecture préalable des chapitre 1 : "Analyse spatiale descriptive", 2 : "Codifier la structure de voisinage" et 3 : "Indices d'autocorrélation spatiale" est recommandée.

Introduction

Les relations entre les valeurs observées sur des territoires proches préoccupent depuis longtemps les géographes. Waldo Tobler a résumé cette problématique par une formule souvent qualifiée de première loi de la géographie : "Tout interagit avec tout, mais deux objets proches ont plus de chance de le faire que deux objets éloignés". La disponibilité de données localisées, associée à des procédures de statistique spatiale désormais pré-programmées dans plusieurs logiciels statistiques, pose la question de la modélisation de cette proximité dans les études économiques. Une première étape reste bien sûr de caractériser cette proximité à l'aide d'indicateurs descriptifs et à l'aide de tests (FLOCH 2012). Une fois l'autocorrélation spatiale des données détectée vient l'étape de la modélisation dans un cadre multivarié. L'objet de ce document de travail est d'aborder la conduite pratique d'une étude d'économétrie spatiale : quel modèle retenir ? Comment en interpréter les résultats ? Quelles en sont les limites ?

Nous illustrerons notre présentation par l'exemple de la modélisation localisée du taux de chômage à l'aide de quelques variables explicatives décrivant les caractéristiques de la population active, de la structure économique, de l'offre de travail et du voisinage géographique. L'objectif ne sera pas de détailler les résultats d'une étude économique [1] mais d'illustrer les techniques mises en œuvre. Nous rappellerons brièvement la définition d'une matrice de voisinage qui décrit les relations de proximité et les tests de corrélation spatiale (décrits plus en détail dans les chapitres 2 : "Codifier la structure de voisinage" et 4 : "Indices d'autocorrélation spatiale"). Nous détaillerons ensuite la spécification, l'estimation et l'interprétation de modèles d'économétrie spatiale.

Les techniques présentées s'appliquent à des domaines qui dépassent le cadre strictement géographique. Plusieurs types de données interconnectées, *i.e.* pouvant interagir entre elles, existent en effet : des points (individus ou entreprises dont on connaît l'adresse), des données par aires géographiques ou administratives (taux de chômage localisés), des réseaux physiques (routes) ou relationnels (élèves d'une même classe) ou des données continues (*i.e.* qui existent en tout point de l'espace). Ces dernières données sont essentiellement issues de la physique, par exemple la hauteur du sol, la température, la qualité de l'air, etc. et relèvent du domaine de la géostatistique (voir chapitre 5 : "Géostatistique"). Elles peuvent néanmoins servir de variables explicatives dans les modèles présentés dans ce document. Un point important à noter est qu'on considère ici des structures de proximité préexistantes, qui n'évoluent pas ou peu. On ne se pose ainsi pas la question de la caractérisation de la formation ou de l'évolution de ces relations de voisinage. On cherche au contraire à caractériser dans quelle mesure la proximité spatiale (ou relationnelle) influence un résultat, en contrôlant de multiples caractéristiques : le taux de chômage dépend-t-il des régions voisines ? les prix des carburants des stations proches ? la non-réponse à une enquête peut-elle se diffuser spatialement ? Si la majorité des applications ont une dimension géographique (ABREU

1. BLANC et al. 2008 traitent cette question de manière détaillée à l'aide d'un modèle d'économétrie spatiale pour la France, LOTTMANN 2013 pour l'Allemagne.

et al. 2004 pour la convergence des PIB régionaux, OSLAND 2010 pour les déterminants des prix de l'immobilier pour des exemples classiques), les domaines d'application sont ainsi plus vastes avec par exemple la mesure des effets de pairs dans les réseaux sociaux (FAFCHAMPS 2015 pour une synthèse), de la proximité idéologique en science politique (BECK et al. 2006) ou la prise en compte de la proximité entre produits pour étudier les effets de substitution en économie industrielle (SLADE 2005). Au sein de l'Insee, ces méthodes ont été utilisées pour étudier la relation entre les prix immobiliers et les risques industriels (GRISLAIN-LETRÉMY et al. 2013), les changements de lieux d'habitation ou la non-réponse dans l'enquête emploi (LOONIS 2012).

Des outils spécifiques ont été développés pour estimer les modèles d'économétrie spatiale. LESAGE et al. 2009 mettent à disposition des programmes MatLab. *GeoDa* est un logiciel libre d'analyse spatiale proposé dans le cadre d'un projet initié par Anselin en 2003 d'analyses spatiales. Il existe également des packages complémentaires pour Stata. Le logiciel le plus complet pour l'estimation de modèles d'économétrie spatiale reste néanmoins R. Les exemples et les codes seront donc présentés à l'aide de ce logiciel.

La suite est organisée comme suit. Les sections 6.1 et 6.2 présentent les raisons économiques et statistiques de la mise en place de ces modèles. La section 6.3 décrit les étapes de l'estimation d'un modèle d'économétrie spatiale. La section 6.4 traite de points techniques plus avancés. La section 6.5 détaille la mise en œuvre sous R à travers la modélisation du taux de chômage par zone d'emploi avant la conclusion. Les lecteurs intéressés par l'approfondissement de ces méthodes pourront notamment se référer à LESAGE et al. 2009, ARBIA 2014 ou LE GALLO 2002, LE GALLO 2004 pour une présentation en langue française.

6.1 Pourquoi tenir compte de la proximité spatiale, organisationnelle ou sociale ?

6.1.1 Les raisons économiques

L'interaction spatiale, organisationnelle ou sociale des agents économiques est classique en économie. ANSELIN 2002a liste ainsi les termes employés pour nommer ces interactions : effets de voisinage, de pair, interactions stratégiques, copie par mimétisme ou par les normes sociales ("*copy-catting*"), concurrence par comparaison ("*yardstick competition*"), etc. Il met notamment en avant deux situations de concurrence entre firmes justifiant le recours à un modèle spatial ou d'interaction.

Dans le premier cas, la décision d'un agent économique (une entreprise par exemple) dépend de la décision des autres agents (ses concurrents). Prenons l'exemple de firmes qui se font concurrence par les quantités (concurrence à la Cournot). La firme i cherche à maximiser sa fonction de profit $\Pi(q_i, q_{-i}, x_i)$ en tenant compte de la production de ses concurrents q_{-i} et des ses caractéristiques x_i qui déterminent ses coûts. La solution de ce problème de maximisation est une fonction de réaction de la forme $q_i = R(q_{-i}, x_i)$.

Dans le second cas, la décision d'un agent économique dépend d'une ressource rare. En reprenant l'exemple d'une firme industrielle, la fonction de profit s'écrit $\Pi(q_i, s_i, x_i)$ avec s_i une ressource rare (qui peut être naturelle, par exemple de l'uranium, ou non, par exemple un composant électronique fabriqué par une seule firme). La quantité s_i qui sera consommée par la firme dépend alors des quantités consommées par les autres firmes et donc de leur production q_{-i}. On retrouve la fonction de réaction précédente.

Cet exemple met en évidence que le recours à un modèle d'interaction est microfondé et que la notion de voisinage n'est pas forcément spatiale. Selon les secteurs industriels, les concurrents d'une entreprise seront ceux proches en termes de distance (les services à la personne, les supermarchés) ou de produits vendus (Coca-Cola et Pepsi). ANSELIN 2002a souligne que ces deux situations amènent à implémenter un même modèle spatial ou d'interaction. Ils sont équivalents d'un point de

vue observationnel. Les processus générateurs des données (PGD) sont différents mais fournissent les mêmes observations. De simples données en coupe ne permettent pas d'identifier la source de l'interaction (une concurrence stratégique par les quantités ou une concurrence sur les ressources dans notre exemple), mais seulement de confirmer sa présence et d'évaluer sa force. À l'instar de l'économétrie classique, il reste nécessaire de réfléchir aux effets identifiés par le modèle et les données.

De plus, les externalités ou effets de voisinages sont couramment contrôlés à l'aide de variables spatiales du type distance (par exemple au plus proche concurrent), ou d'indicateurs agrégés par zone géographique (par exemple le nombre de concurrents). Ce type de variable peut s'interpréter comme des variables spatialement décalées (*i.e.* fonction des observations dans les zones voisines), avec une définition *a priori* de relations de voisinage. L'économétrie spatiale justifie et généralise ainsi ces choix empiriques.

6.1.2 Les raisons économétriques

Les raisons économétriques renvoient aux insuffisances de la modélisation linéaire classique (et de l'estimation associée par la méthode des Moindres Carrés Ordinaire -MCO-) lorsque les hypothèses nécessaires à sa mise en œuvre ne sont plus vérifiées. LESAGE et al. 2009 présentent ainsi plusieurs arguments techniques justifiant l'emploi de méthodes spatiales. On observe fréquemment avec des données spatiales une autocorrélation spatiale des résidus, *i.e.* une dépendance entre des observations proches. Cette dépendance des observations peut se traduire soit par une perte d'efficacité des MCO (les estimateurs seront sans biais mais moins précis, et les tests n'auront plus les propriétés statistiques usuelles), soit par des estimateurs biaisés. Si le modèle omet une variable explicative spatialement corrélée à la variable d'intérêt, il y a ainsi biais de variable omise. De plus, la confrontation de plusieurs modèles d'économétrie spatiale permet de discuter l'incertitude du processus générateur des données, qui n'est jamais connu, et de vérifier ainsi la robustesse des résultats.

Les raisons économétriques de recourir aux modèles spatiaux sont nombreuses, dans la mesure où les analyses descriptives mettent en évidence des effets de proximité et des corrélations spatiales. Dans les études appliquées, il est parfois difficile de lier les aspects économétriques et économiques justifiant de la prise en compte de la dépendance spatiale et les causalités de nature économique sont difficiles à établir à partir de modèles économétriques spatiaux (GIBBONS et al. 2012).

6.2 Autocorrélation, hétérogénéité, pondérations : quelques rappels de statistique spatiale

6.2.1 La nature des effets spatiaux dans les modèles de régression

La célèbre phrase de Waldo Tobler, citée en introduction, résume bien les choses, mais les simplifie sans doute un peu. ANSELIN et al. 1988, distinguent l'autocorrélation (la dépendance spatiale) et l'hétérogénéité (la non-stationnarité spatiale). Divers phénomènes, de mesure (choix du découpage territorial), d'externalités ou de débordement ("*spillover*") peuvent conduire à rendre les observations (variable endogène, exogène ou terme d'erreur) dépendantes spatialement. Il y a alors autocorrélation (positive) lorsqu'il y a similarité entre les valeurs observées et leur localisation. Ce chapitre traite principalement des méthodes de prise en compte de cette corrélation spatiale dans les modèles de régression, détaillés en section 6.3. L'hétérogénéité spatiale renvoie quant à elle à des phénomènes d'instabilité structurelle dans l'espace. Cette autre forme de prise en compte de l'espace est détaillée dans le chapitre 9 : "Régression géographiquement pondérée". Elle part de l'idée que les variables explicatives peuvent être les mêmes mais ne pas avoir le même effet en tout point. Les paramètres du modèle sont alors variables. Le terme d'erreur peut être différent selon la zone géographique. On parle alors d'hétérogénéité spatiale. Par exemple, pour définir l'indice des

prix de l'immobilier ancien Insee-Notaires, environ 300 strates sont définies selon la nature du bien (appartement ou maison) et la zone géographique. Le prix du m^2, d'une pièce complémentaire ou d'une autre caractéristique est en effet supposé différent selon ces différentes strates. Le marché est segmenté.

Ce partage "pédagogique" entre autocorrélation et hétérogénéité ne doit pas faire oublier les interactions entre les deux (ANSELIN et al. 1988 ; LE GALLO 2002 ; LE GALLO 2004). Il n'est pas toujours facile de distinguer chacune des deux composantes, et la mauvaise spécification de l'une peut être la cause de l'autre. Les tests classiques de l'hétéroscédasticité (*i.e.* une forme particulière d'hétérogénéité sur le terme d'erreur) sont affectés par l'autocorrélation spatiale, et inversement les tests d'autocorrélation spatiale le sont par l'hétéroscédasticité. Il n'y a pas de solution simple pour intégrer simultanément ces deux phénomènes, en dehors du simple ajout d'indicatrices de territoires dans les modèles d'autocorrélation. De plus, la corrélation des valeurs observées fait que l'information apportée par les données est moins riche que dans le cas où les données sont indépendantes. En cas d'autocorrélation, on observe une seule réalisation du processus générateur des données. Tout ceci plaide pour une approche exploratoire préalable des données. Selon la question, la méthodologie traitera en premier lieu l'autocorrélation spatiale des observations (*i.e.* les liens entre les unités proches) ou l'hétérogénéité des comportements (*i.e.* leur variabilité selon la localisation).

6.2.2 La matrice des poids

Pour mesurer la corrélation spatiale entre agents ou zones géographiques, tout commence par la définition *a priori* des relations de voisinage entre les agents ou les zones géographiques. Ces relations ne peuvent pas être estimées par le modèle. Si nous observons N régions, il y a $N(N-1)/2$ couples différents de régions. Il n'est donc pas possible d'identifier des relations de corrélation entre ces N régions sans faire des hypothèses sur la structure de cette corrélation spatiale. Pour N agents ou zones géographiques, cela revient à définir une matrice carrée de taille $N \times N$, dite matrice de voisinage et notée W, dont les éléments diagonaux sont nuls (on ne peut pas être son propre voisin). La valeur des éléments non diagonaux est le fruit de l'expertise. De nombreuses matrices de voisinage ont été proposées dans la littérature. Leur construction avec le logiciel R est détaillée dans le chapitre 2 : "Codifier la structure de voisinage".

6.2.3 Les méthodes exploratoires

Avant de spécifier un modèle d'économétrie spatiale, il convient de vérifier qu'il y a bien un phénomène spatial à prendre en compte. Cela commence par une caractérisation de l'autocorrélation spatiale à l'aide de représentations graphiques (carte) et de tests statistiques décrits dans le chapitre 3 : "Indices d'autocorrélation spatiale".

Le principal indicateur[2] est celui de Moran qui mesure l'association globale : $I = \frac{N}{\sum_i \sum_j w_{ij}} \cdot \frac{\sum_i \sum_j w_{ij}(y_i-\bar{y})(y_j-\bar{y})}{\sum_i (y_i-\bar{y})^2}$, avec w_{ij} le poids correspondant au coefficient situé sur la i-ème ligne et la j-ème colonne de la matrice de voisinage W. Les bornes de l'indicateur de Moran I sont comprises entre -1 et 1 et dépendent de la matrice de poids utilisée. La borne supérieure est notamment égale à 1 si la matrice est standardisée en ligne, la borne inférieure reste différente en toute généralité de -1. Une corrélation positive signifie que les zones avec de hautes ou de basses valeurs pour y se regroupent, une corrélation négative que des zones géographiques proches ont des valeurs de y très différentes. Sous l'hypothèse H_0 d'absence d'autocorrélation spatiale ($I = 0$), la

2. Les indicateurs de Geary et de Getis et Ord, ainsi que les autres indicateurs locaux, sont présentés dans FLOCH 2012.

statistique $I^* = \frac{I-\mathbb{E}(I)}{\sqrt{\mathbb{V}(I)}}$ suit asymptotiquement une loi normale $\mathcal{N}(0,1)$. Rejeter l'hypothèse nulle du test de Moran revient donc à conclure à la présence d'autocorrélation spatiale. Ce test reste bien sûr dépendant du choix de la matrice de voisinage W. De plus, le rejet de H_0 ne signifie pas qu'un modèle d'économétrie spatiale soit nécessaire mais que celui-ci doive être envisagé. Il peut en effet ne refléter que la répartition spatiale d'une variable sous-jacente. Par exemple, si le modèle sous-jacent est $Y = X \cdot \beta + \varepsilon$ avec β un paramètre à estimer, $\varepsilon \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma^2\right)$ et X une variable autocorrélée spatialement, un test de Moran conclura à l'autocorrélation spatiale de la variable Y. Pour autant, le modèle linéaire liant Y et X n'est pas un modèle spatial, et peut être estimé classiquement à l'aide des MCO.

Des indicateurs locaux (par zone géographique i, dits LISA pour *Local Indicators of Spatial Association*) ont été définis pour mesurer la propension d'une zone à regrouper de fortes ou faibles valeurs de y ou au contraire des valeurs très diverses. Leur calcul est détaillé dans le chapitre 3 : "Indices d'autocorrélation spatiale".

6.3 Estimer un modèle d'économétrie spatiale

6.3.1 La galaxie des modèles d'économétrie spatiale

ELHORST 2010 a établi une classification des principaux modèles d'économétrie spatiale, en s'appuyant sur les trois types d'interaction spatiale issus du modèle fondateur de MANSKI 1993 :

— une interaction endogène, lorsque la décision économique d'un agent ou d'une zone géographique va dépendre de la décision de ses voisins ;
— une interaction exogène, lorsque la décision économique d'un agent va dépendre des caractéristiques observables de ses voisins ;
— une corrélation spatiale des effets liée à de mêmes caractéristiques inobservées.

Ce modèle s'écrit sous forme matricielle [3] :

$$\begin{aligned} Y &= \rho \cdot WY + X \cdot \beta + WX \cdot \theta + u \\ u &= \lambda \cdot Wu + \varepsilon \end{aligned} \tag{6.1}$$

Avec les paramètres β pour les variables explicatives exogènes, ρ pour l'effet d'interaction endogène (de dimension 1) dit autorégressif spatial, θ pour les effets d'interaction exogène (de dimension égale au nombre de variables exogènes K) et λ pour l'effet de corrélation spatiale des erreurs dit autocorrélation spatiale. Dans la suite du document, nous emploierons le terme de corrélation spatiale pour désigner un de ces 3 types d'interaction spatiale.

Le modèle de MANSKI 1993 n'est pas identifiable sous cette forme, c'est-à-dire qu'on ne peut pas estimer à la fois β, ρ, θ, et λ. Prenons son exemple des effets de pairs pour en donner l'intuition. Supposons que les mauvais résultats scolaires d'une classe s'expliquent par la composition sociale

3. Par souci de simplification, la constante du modèle est ici incluse dans la matrice des variables explicatives X. Dans le cas d'une matrice de contiguïté, $W \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ représente le nombre de voisins de chaque observation. Si ce nombre de voisins est le même pour tous les individus, la constante β_0 et le terme $W \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \cdot \theta_0$ ne sont pas identifiables séparément. De plus, le nombre de voisins (ou le nombre moyen si la matrice de voisinage est normée par ligne) n'a pas forcément un sens économique clair. C'est pourquoi on trouve dans la littérature une présentation des modèles où la constante n'est pas incluse dans la matrice des variables explicatives X.

de la classe (interaction exogène) et le fait d'avoir de mauvais professeurs (caractéristique inobservée). On constatera alors une forte corrélation des résultats des élèves au sein de la classe mais cela ne signifie pas que le fait d'être avec des élèves d'un niveau scolaire plus faible (interaction endogène) a un effet.

Une première solution, pour rendre le modèle identifiable, est de supposer que les matrices de voisinage W ne sont pas identiques pour les trois interactions spatiales. Il y aurait par exemple des relations de voisinage définies par W_ρ pour le paramètre autorégressif et W_λ pour l'autocorrélation spatiale. SLADE 2005 définit ainsi deux matrices de voisinage distinctes pour étudier les effets prix en économie industrielle : W_ρ étant fonction de la distance entre entreprises concurrentes et W_X d'un indicateur de proximité entre les produits vendus. Une autre solution consiste à supprimer l'une des 3 formes de corrélation spatiale, représentées par les paramètres ρ, θ et λ. C'est la solution privilégiée dans la littérature empirique.

La matrice de voisinage doit respecter plusieurs contraintes techniques (LEE 2004 ; ELHORST 2010) pour assurer notamment le caractère inversible des matrices $I-\rho W$ et $I-\lambda W$, et l'identification des modèles. On peut retenir que les matrices usuelles de contiguïté ou de distance inverse respectent ces contraintes. Ce n'est pas forcément le cas de matrices "atypiques" créées par exemple pour les relations de proximité sociale. Il n'est par exemple pas possible d'avoir uniquement des îles (une zone qui n'a pas de voisin) ou au contraire que tout le monde soit le voisin de tout le monde. On doit de plus supposer que $|\rho|<1$ et $|\lambda|<1$ (critères qu'on peut intuitivement rapprocher des conditions de stationnarité pour les solutions d'un modèle de type ARMA).

Trois principaux types de modèles peuvent être déduits du modèle de MANSKI 1993 selon la contrainte utilisée, $\theta=0$, $\lambda=0$ ou $\rho=0$.
Le cas $\rho=0$ (modèle SDEM, *Spatial Durbin Error Model*) peut être envisagé si on suppose qu'il n'y a pas d'interaction endogène et que l'accent est mis sur les externalités de voisinage. Ce modèle reste néanmoins d'un usage moins courant (LESAGE 2014).

Si on suppose que le modèle est tel que $\theta=0$, on trouve le modèle de Kelejian-Prucha (ou également nommé SAC, *Spatial Autoregressive Confused*, KELEJIAN et al. 2010a pour le modèle hétéroscédastique) :

$$\begin{aligned} Y &= \rho \cdot WY + X \cdot \beta + u \\ u &= \lambda \cdot Wu + \varepsilon \end{aligned} \tag{6.2}$$

Les estimateurs de β du modèle de Kelejian-Prucha présentent le défaut d'être biaisés et non convergents dans le cas où le vrai modèle inclut des interactions exogènes WX (LESAGE et al. 2009). Il y a en effet dans ce cas biais de variables omises. De plus, LE GALLO 2002 souligne que choisir une même matrice de voisinage W pour ce modèle engendre une identification faible des paramètres.

Au contraire, si on suppose que le modèle est tel que $\lambda=0$, $Y=\rho\cdot WY+X\cdot\beta+WX\cdot\theta+\varepsilon$, dit modèle spatial de Durbin (SDM, *Spatial Durbin Model*) , alors les estimateurs seront non biaisés (et les statistiques de test valides) même si, en réalité, nous sommes en présence d'erreurs autocorrélées spatialement (SEM). Ce modèle est ainsi plus robuste à un mauvais choix de spécification.

Ces deux modèles (Kelejian-Prucha et SDM) incluent les cas particuliers du modèle spatial autorégressif (SAR, *Spatial AutoRegression*) : $Y=\rho\cdot WY+X\cdot\beta+\varepsilon$) et du modèle à erreurs autocorrélées spatialement (SEM, *Spatial Error Model*) : $Y=X\cdot\beta+u$ et $u=\lambda\cdot Wu+\varepsilon$). Pour obtenir ce dernier modèle à partir du modèle spatial de Durbin, on pose $\theta=-\rho\beta$ (hypothèse dite de facteur commun). Le modèle SDM s'écrit dans ce cas : $Y=X\cdot\beta+\rho\cdot W(Y-X\cdot\beta)+\varepsilon$. En notant $u=Y-X\cdot\beta$, on retrouve bien le modèle SEM. Le modèle à interactions exogènes (noté SLX, *Spatial Lag X*) correspond au cas $\lambda=\rho=0$ et $\theta\neq 0$.

Il existe par ailleurs des versions plus générales de ces modèles, qui autorisent une variation des effets de voisinage selon l'ordre de voisinage ou selon les interactions prises en compte. Ils correspondent à des versions spatiales des modèles temporels *ARMA(p,q)*.

Dans le cadre d'une étude économique, on ne présente pas l'ensemble de ces modèles. Les critères statistiques et la cohérence avec la question économique permettent de retenir une spécification plutôt qu'une autre.

6.3.2 Critères statistiques du choix de modèle

Deux approches principales ont été utilisées pour le choix des modèles. Ces approches "pratiques" reposent sur l'hypothèse que la matrice de voisinage soit connue et que les variables explicatives soient exogènes. Sous l'hypothèse de normalité des résidus $\varepsilon \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma^2\right)$, elles reposent sur une estimation par maximum de vraisemblance des modèles et les tests statistiques associés [4]. La première dite "approche ascendante" ou *bottom-up* (figure 6.1) consiste à commencer avec le modèle non spatial (LE GALLO 2002 pour une synthèse). Des tests du multiplicateur de Lagrange (ANSELIN et al. 1996 pour des tests de spécification des modèles SAR et SEM, robustes à la présence d'autres types d'interactions spatiales) permettent ensuite de trancher entre le modèle SAR, SEM ou le modèle non spatial. Cette approche a été celle plébiscitée jusqu'aux années 2000 car les tests développés par ANSELIN et al. 1996 s'appuient sur les résidus du modèle non spatial. Ils sont donc peu coûteux d'un point de vue computationnel. FLORAX et al. 2003 ont également montré, à l'aide de simulations, que cette procédure était la plus performante dans le cas où le vrai modèle est un modèle SAR ou SEM.

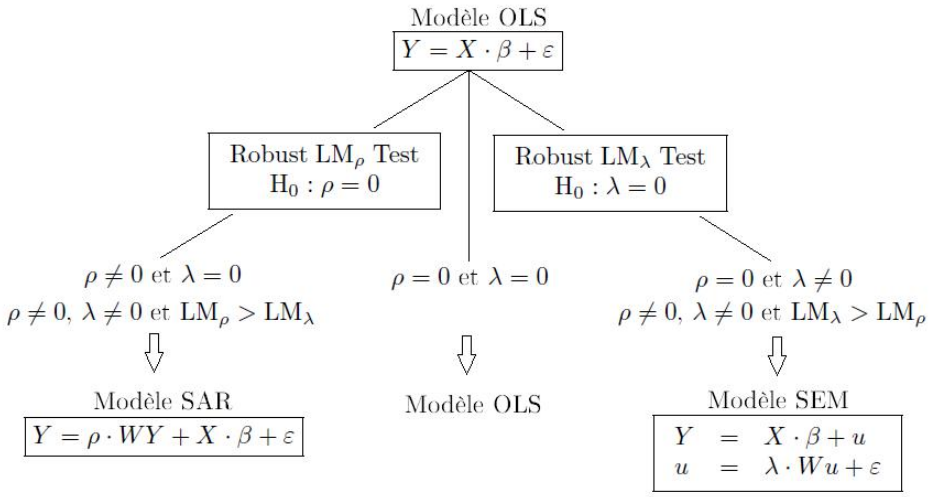


FIGURE 6.1 – Approche *bottom-up*
Source : FLORAX *et al. 2003*

La deuxième approche dite "approche descendante" ou *top-down* (figure 6.2) consiste à commencer avec le modèle spatial de Durbin. À partir des tests du rapport de vraisemblance, on en déduit le modèle le plus adapté aux observations. L'amélioration des performances informatiques

4. D'autres méthodes d'estimation existent. Dans le cas de variables explicatives endogènes, FINGLETON et al. 2008 FINGLETON et al. 2012 proposent une estimation par variables instrumentales et la méthode des moments généralisée. LESAGE et al. 2009 proposent une estimation bayésienne. Enfin, pour relâcher le cadre paramétrique, LEE 2004 propose une estimation par quasi maximum de vraisemblance.

a permis de rendre aisée l'estimation de ces modèles plus complexes, dont le modèle spatial de Durbin pris comme référence dans le livre de LESAGE et al. 2009.

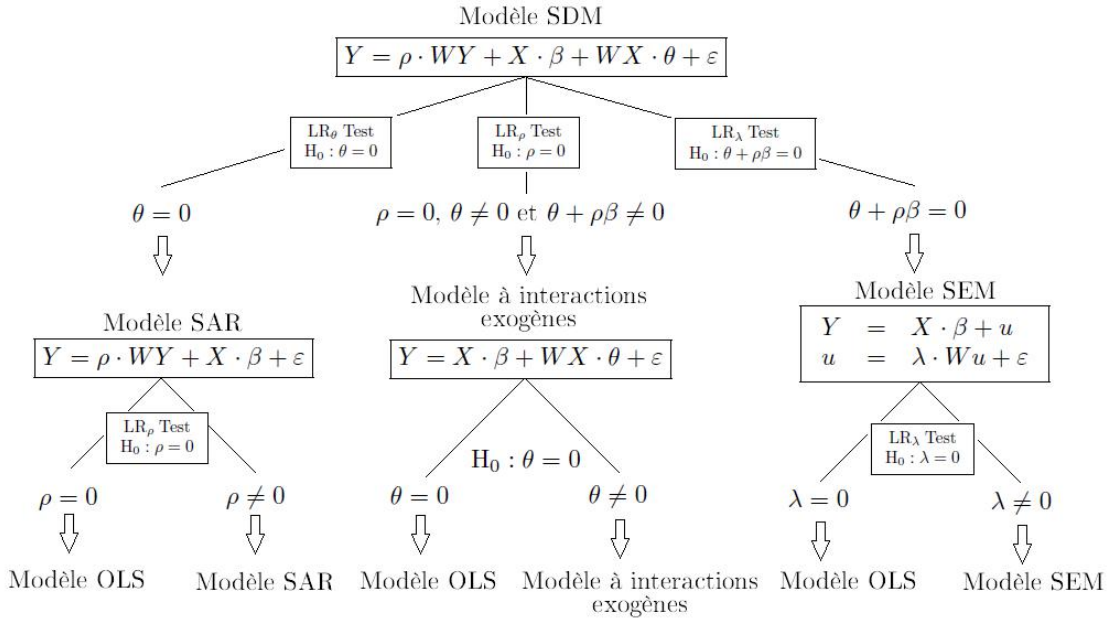


FIGURE 6.2 – Approche *top-down*
Source : LESAGE *et al. 2009*

ELHORST 2010 propose une approche "mixte" représentée en figure 6.3. Elle consiste à commencer par l'approche ascendante mais, en cas d'interactions spatiales ($\rho \neq 0$ ou $\lambda \neq 0$), au lieu de choisir directement un modèle SAR ou SEM, à étudier le modèle spatial de Durbin. Cela permet de confirmer à l'aide de plusieurs tests (multiplicateur de Lagrange, rapport de vraisemblance) la pertinence du modèle choisi. Cela permet également d'intégrer les interactions exogènes dans l'analyse. Enfin, en cas d'incertitude, c'est le modèle *a priori* le plus robuste (le modèle spatial de Durbin) qui est choisi. Prenons le cas où, à partir des résidus du modèle OLS, les tests du multiplicateur de Lagrange (LM_ρ et LM_λ)[5] concluent à la présence d'un terme autorégressif, *i.e.* $\rho \neq 0$ et $\lambda = 0$ (branche de gauche de la figure 6.1). On estime alors le modèle SDM. À l'aide d'un test du rapport de vraisemblance ($\theta = 0$), on peut alors choisir entre le modèle SAR et le modèle SDM. Dans le cas où les tests concluent à la présence d'autocorrélation résiduelle, *i.e.* $\rho = 0$ et $\lambda \neq 0$ (branche de droite de la figure 6.2), on se ramène au modèle SDM ($\rho \neq 0$ et $\theta \neq 0$), puis un test du rapport de vraisemblance de l'hypothèse de facteur commun ($\theta = -\rho\beta$) permet de choisir entre le modèle SEM et le modèle SDM. Dans le cas où les tests soulignent l'absence de corrélation spatiale, *i.e.* $\rho = 0$ et $\lambda = 0$, le modèle à interactions exogènes (SLX) est estimé. Des tests du rapport de vraisemblance permettent de choisir entre les modèles OLS, SLX et SDM. Enfin, dans le cas où les tests concluent à la présence simultanée de corrélation endogène et résiduelle, *i.e.* $\rho \neq 0$ et $\lambda \neq 0$, le modèle SDM est estimé.

La matrice de voisinage W a pour dimension le carré du nombre d'observations. Or le calcul de la vraisemblance de ces modèles spatiaux fait notamment intervenir des déterminants incluant cette matrice. Le coût computationnel peut donc être important lorsque le nombre d'observations devient élevé. LESAGE et al. 2009 consacrent ainsi un chapitre aux enjeux computationnels (et

5. Il existe deux versions de ces tests, l'une robuste à la présence d'autres formes de corrélation spatiale, l'autre non (ANSELIN et al. 1996).

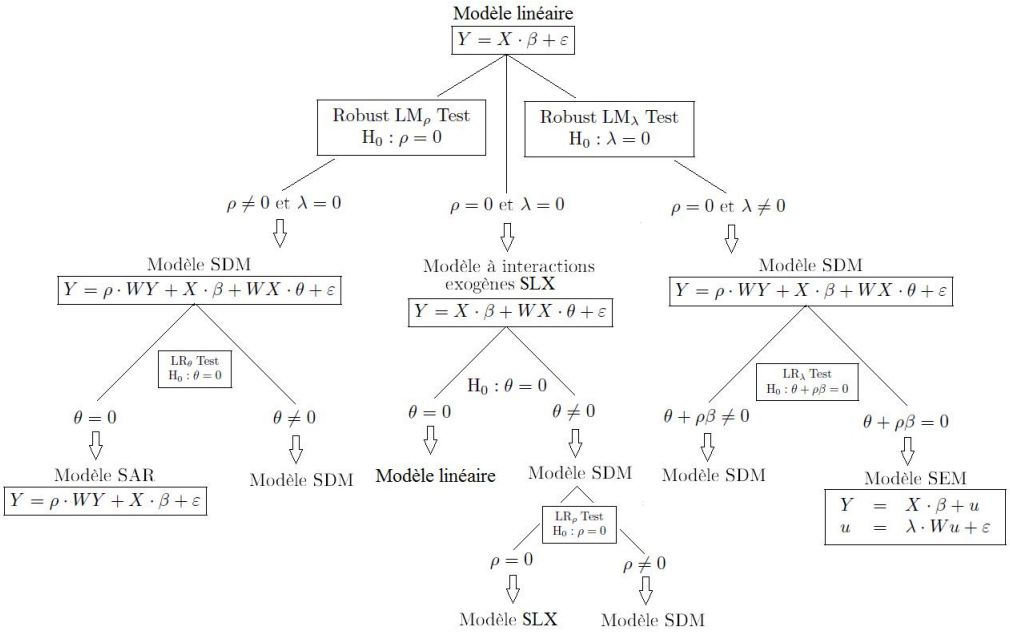


FIGURE 6.3 – Approche d'ELHORST 2010 pour le choix d'un modèle d'économétrie spatiale
Source : ELHORST 2010.

aux méthodes pour les résoudre) associés à l'estimation de ces modèles. En pratique, le nombre d'observations est souvent limité à quelques milliers.

Ces règles ne doivent pas être considérées comme intangibles [6], mais plutôt comme de bonnes pratiques. Il ne sert en effet à rien d'estimer directement un modèle SAR, complexe à interpréter, si ni l'analyse économique, ni l'analyse statistique ne le justifient.

6.3.3 L'interprétation des résultats : attention aux rétroactions

L'économétrie spatiale s'écarte du cadre habituel des modèles linéaires lorsque des variables spatialement décalées WY sont présentes dans le modèle. L'interprétation classique des modèles linéaires reste en revanche valide si seule une autocorrélation spatiale des erreurs est prise en compte (modèle SEM).

En présence d'une variable spatialement décalée WY les paramètres associés aux variables explicatives ne s'interprètent pas comme dans le cadre habituel des modèles linéaires. En effet, du fait des interactions spatiales, la variation d'une variable explicative pour une zone donnée affecte directement son résultat et indirectement les résultats de toutes les autres zones. Les paramètres estimés interviennent alors dans le calcul d'un effet multiplicateur qui est global car il affecte l'ensemble de l'échantillon.

En revanche, l'interprétation des paramètres associés aux variables explicatives reste identique lorsque le modèle ne comporte qu'une autocorrélation des erreurs (modèle SEM). Dans ce cas, il existe un effet de diffusion global lié aux erreurs autocorrélées spatialement : la variation d'une variable explicative pour une zone donnée affecte directement son résultat et indirectement les résultats de toutes les autres zones, mais sans que la valeur de cet effet soit démultipliée.

Lorsqu'on considère des modèles avec variables explicatives spatialement décalées (SLX) les paramètres associés aux variables explicatives permettent de calculer un effet local dans la mesure

6. L'approche séquentielle des tests peut de plus engendrer un biais car la zone de rejet des tests du rapport de vraisemblance (LR) devrait en théorie tenir compte des tests préalables du multiplicateur de Lagrange (LM).

où la variation d'une variable explicative affecte directement son résultat et indirectement le résultat des zones voisines, mais pas celui des zones voisines de ces voisines.

Pour formaliser les différents impacts, nous reprenons le cadre défini par LESAGE et al. 2009.

Le modèle SAR est $Y = \rho \cdot WY + X\beta + \varepsilon$. Il peut se réécrire de plusieurs manières, en notant r l'indice pour une variable explicative et S_r des matrices carrées de la taille du nombre d'observations :

$$\begin{aligned} Y &= (1-\rho W)^{-1} X\beta + (1-\rho W)^{-1} \varepsilon \\ &= \sum_{r=1}^{k} (1-\rho W)^{-1} \beta_r X_r + (1-\rho W)^{-1} \varepsilon \\ &= \sum_{r=1}^{k} S_r(W) X_r + (1-\rho W)^{-1} \varepsilon \end{aligned} \tag{6.3}$$

$$\text{Avec } Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \text{ et } S_r(W) = \begin{pmatrix} S_r(W)_{11} & S_r(W)_{12} & \cdots & S_r(W)_{1n} \\ S_r(W)_{21} & S_r(W)_{22} & & \\ \vdots & \vdots & \ddots & \\ S_r(W)_{n1} & S_r(W)_{n2} & \cdots & S_r(W)_{nn} \end{pmatrix}$$

La valeur prédite est donc $\hat{y} = (1-\hat{\rho}W)^{-1} X\hat{\beta}$ [7] et non $X\hat{\beta}$ comme dans un modèle linéaire classique.

On a de plus $\mathbb{E}(y) = (1-\rho W)^{-1} X\beta$. L'effet marginal (pour une variable quantitative) d'une modification de la variable X_r pour l'individu i n'est pas β_r mais $S_r(W)_{ii}$, la valeur diagonale de rang i de la matrice S_r. À la différence des séries temporelles où il n'y a qu'une direction à prendre en compte (y_t dépend de y_{t-1} qui n'est expliquée que par des valeurs passées), l'économétrie spatiale est multidirectionnelle. Une modification de mon territoire impacte mes voisins, ce qui m'impacte en retour. Il faut en tenir compte pour l'analyse globale des résultats.

Par ailleurs, l'effet marginal apparaît différent pour chaque zone [8]. Les termes diagonaux de la matrice S_r sont les effets directs, pour chaque zone, d'une modification de la variable X_r dans la même zone. Les autres termes représentent des effets indirects, *i.e.* l'impact de la modification de la variable X_r dans une zone sur une autre zone. Pour l'ensemble des zones (niveau global) on peut donc calculer des effets directs et indirects obtenus en faisant la moyenne de ces effets (LESAGE et al. 2009) :

— L'effet direct moyen correspond à la moyenne des termes diagonaux de la matrice S_r, *i.e.* $\frac{1}{n}\text{trace}(S_r)$. L'interprétation de cet indicateur se rapproche de celle des coefficients β d'un modèle linéaire non spatial calculés par la méthode des MCO.

— L'effet total moyen correspond à une moyenne de l'ensemble des termes de la matrice S_r, $\frac{1}{n}\sum_i [\sum_k S_r(W)_{ik}]$. Il peut s'interpréter de deux manières, soit comme la moyenne des n effets sur une zone i d'une modification d'une unité de la variable X_r dans toutes les zones, *i.e.* $\sum_k S_r(W)_{ik}$ (la somme des termes en ligne de la matrice S_r), soit comme la moyenne des n effets d'une modification d'une unité de la variable X_r dans une zone i sur l'ensemble des zones, *i.e.* $\sum_k S_r(W)_{ki}$ (la somme des termes en colonne de la matrice S_r).

— L'effet indirect moyen est la différence entre l'effet total moyen et l'effet direct moyen.

7. Ce n'est pas la prédiction optimale, voir THOMAS-AGNAN et al. 2014 pour la prédiction optimale d'un modèle SAR.

8. On retrouve cette caractéristique pour l'effet marginal d'un modèle Probit par exemple. Le modèle est $\mathbb{E}(Y|X) = \mathbb{P}(Y=1|X) = \Phi(\beta X)$ avec Φ la fonction de répartition d'une loi normale centrée-réduite. L'effet marginal d'une variable X_r est alors $\beta_r \cdot \varphi(\beta X)$ et diffère donc pour chaque individu. Une solution est alors d'estimer l'effet marginal moyen $\beta_r \cdot \overline{\varphi(\beta X)}$.

Les indicateurs sont identiques pour le modèle de Kelejian-Prucha. De tels indicateurs peuvent être définis pour le modèle SDM $Y = \rho \cdot WY + X\beta + WX \cdot \theta + \varepsilon$, mais leurs calculs doivent tenir compte des interactions exogènes $WX \cdot \theta$. La matrice $S_r(W)$ s'écrit en effet dans ce cas $(1-\rho W)^{-1}(I_n\beta_r + W\theta_r)$, au lieu de $(1-\rho W)^{-1}\beta_r$ dans le cas du modèle SAR.

Lorsqu'une interaction exogène $WX \cdot \theta$ est présente mais qu'il n'y a pas d'interaction endogène (modèles SLX et SDEM), l'effet direct d'une variable X_r est β_r, l'effet indirect est θ_r.

Dans tous les cas, le calcul de la précision de ces estimateurs est complexe. Pour ce calcul, LESAGE et al. 2009 s'appuient ainsi sur des simulations bayésiennes de Monte-Carlo par Chaîne de Markov (MCMC) [9].

Par ailleurs, ces effets dépendent en premier lieu du voisinage proche. Pour le modèle SAR, on peut noter que l'effet direct moyen est supérieur en valeur absolue à l'effet marginal du modèle linéaire non spatial, $|S_r| > |\beta_r|$. Les termes diagonaux de la matrice de voisinage W sont en effet nuls. La décomposition en séries entières $(1-\rho W)^{-1} = \left(I_n + \rho W + \rho^2 W^2 + \cdots\right)$ montre que le premier terme de rétroaction (et qui domine les autres termes d'ordre supérieur) est proportionnel à ρ^2. L'analyse des effets par ordre de voisinage (distinguer l'effet direct, l'effet des voisins, des voisins des voisins, etc.) est également developpée par LESAGE et al. 2009.

En conclusion, pour l'interprétation globale d'un modèle avec interaction endogène, il est utile de calculer pour chaque variable, l'effet direct moyen ($\frac{1}{n}\text{trace}(S_r)$) et l'effet indirect moyen ($\frac{1}{n}\left[\sum_j \sum_k S_r(W)_{kj} - \text{trace}(S_r)\right]$). Calculer l'effet induit par l'espace ($\frac{1}{n}\text{trace}(S_r) - \hat{\beta}_r$) permet également d'illustrer la force des effets de rétroaction.

6.4 Limites et difficultés économétriques

6.4.1 Que faire des données manquantes ?

En économétrie classique, on observe un échantillon de n individus. Si quelques individus présentent des valeurs manquantes, ils sont généralement exclus de l'analyse. En l'absence de sélection liée à la non-réponse (le processus de non-réponse est indépendant des variables de notre modèle), cela réduit la taille de l'échantillon mais n'empêche pas la mise en œuvre des méthodes économétriques.

En économétrie spatiale, on observe une seule réalisation du processus générateur des données (une analogie peut être effectuée avec les séries temporelles, les paramètres d'un modèle ARMA étant estimés à l'aide d'une seule trajectoire temporelle). Si l'observation de la distribution spatiale est incomplète (il y a des valeurs manquantes), il n'est pas possible d'estimer le modèle. Une solution consiste à interpoler les valeurs manquantes à l'aide de techniques de géostatistique (ANSELIN 2001), mais cela a pour incidence de mesurer les variables avec erreurs [10], ou d'utiliser une estimation adaptée (par exemple algorithme EM espérance-maximisation, WANG et al. 2013b pour le modèle SAR). Ces solutions ne sont néanmoins possibles que pour un faible pourcentage de valeurs manquantes.

Une autre implication est qu'il n'est pas aisé de mettre en place ces techniques sur données individuelles d'enquête. Dans le cas général, l'économétrie spatiale n'est pas adaptée aux données

9. Les méthodes de Monte-Carlo par Chaîne de Markov sont des algorithmes d'échantillonnage permettant de générer des échantillons d'une loi de probabilité complexe (pour en déduire par exemple la précision d'une statistique). Elles s'appuient sur un cadre bayésien et une chaîne de Markov dont la loi limite est la distribution à échantillonner.

10. L'interpolation peut également être utile lorsque les niveaux géographiques servant à mesurer la variable à expliquer et les variables explicatives sont différentes, par exemple les prix de logements connus au niveau de l'adresse ou de la commune et des indicateurs de pollution atmosphérique mesurés à l'aide de capteurs dont les localisations diffèrent.

d'enquêtes. En effet, dans ce cas on n'observe que des relations de voisinage partielles, pour les seuls individus enquêtés. Il faut alors faire l'hypothèse complémentaire et très forte que les observations des voisins non enquêtés sont exogènes, *i.e.* qu'elles ne modifient pas les effets de voisinage pour les seuls individus enquêtés. Dans le chapitre 11 "Économétrie spatiale sur données d'enquêtes", Lardeux et Merly-Alpa montrent qu'il n'est possible de détecter la corrélation spatiale générée par un modèle SAR uniquement pour un plan de sondage par grappes géographiques. Avec de faibles taux de sondage et des plans de sondages classiques (stratifiés ou systématiques), seuls les effets directs peuvent être estimés. Ce point est développé dans le chapitre 11 : "économétrie spatiale sur données d'enquête".

6.4.2 Le choix de la matrice de poids

Pour définir une matrice de voisinage, les contraintes sont fortes, puisque l'on recherche une description simple (afin que le modèle soit identifiable), mais adéquate des relations entre territoires. De nombreux auteurs soulignent la sensibilité des résultats au choix de cette matrice (CORRADO et al. 2012 ; HARRIS et al. 2011), alors que LESAGE et al. 2009 considèrent que ces conclusions proviennent d'une mauvaise interprétation des modèles et que cette sensibilité supposée à la matrice de poids est "le plus grand mythe" de l'économétrie spatiale. Les effets directs et indirects seraient plus robustes au choix de W que les estimateurs des paramètres, qui n'ont eux pas d'interprétation immédiate. On peut néanmoins souscrire à la remarque de HARRIS et al. 2011 : "L'économétrie spatiale souligne l'importance du choix de la matrice W mais nous renseigne peu sur les critères pour effectuer ce choix", difficultés qui ont contribué au scepticisme de plusieurs économistes (GIBBONS et al. 2012). Ces considérations montrent la complexité de la détermination de la matrice W qui reste un sujet de controverses scientifiques.

On a vu que les modèles traitent en général la matrice W comme exogène. D'autres méthodes s'appuient néanmoins sur les données utilisées pour déterminer la matrice des poids. ALDSTADT et al. 2006 définissent ainsi un algorithme de construction de la matrice W à partir des indicateurs locaux d'autocorrélation spatiale des variables d'intérêt. Il est également possible d'estimer les poids à partir de modèles économétriques avec des contraintes fonctionnelles *a priori* faibles (BHATTACHARJEE et al. 2013). Ces dernières approches sont souvent lourdes en calcul et plus difficiles à implémenter. De plus, une description plus réaliste et plus conforme à la réalité économique risque d'introduire de l'endogénéité. Des travaux faisant intervenir des matrices endogènes ont été récemment proposés (KELEJIAN et al. 2014).

Dernier point, la matrice W est considérée fixe, ce qui contraint le cadre de l'analyse économique. Par exemple, dans le cas de matrice de voisinage mesurant la distance entre entreprises ou produits, WAELBROECK 2005 souligne que l'arrivée (ou le départ) d'une entreprise ou d'un produit est un évènement endogène qui devrait amener à modifier les relations de voisinages, ce que ne permet pas la méthodologie usuelle.

6.4.3 Et si le phénomène est hétérogène spatialement ?

Deux formes d'hétérogénéité existent.

La première est l'hétéroscédasticité. Les paramètres du modèle sont les mêmes mais pas sa variabilité individuelle. Une autocorrélation spatiale des erreurs $(I-\lambda W)^{-1}\varepsilon$ (modèle SEM) peut s'interpréter comme un effet aléatoire spatial (on suppose que les effets individuels au sein d'un voisinage sont proches, faute de pouvoir estimer des effets fixes) et donc une forme particulière d'hétéroscédasticité et de corrélation spatiale (LESAGE et al. 2009). Une solution alternative à un modèle d'économétrie spatiale serait de définir la forme de l'hétéroscédasticité et de la corrélation spatiale de la matrice de variances-covariances (DUBIN 1998), de définir des clusters spatiaux (BARRIOS et al. 2012) ou d'adopter une correction spatiale du type Newey-West (FLACHAIRE 2005). Enfin, des développements récents de l'économétrie spatiale

relâchent l'hypothèse d'homoscédasticité des résidus ε des modèles présentés dans cette introduction. KELEJIAN et al. 2007 KELEJIAN et al. 2010b ont ainsi proposé une méthode paramétrique de type HAC (*Heteroscedasticity and Autocorrelation Consistent*), issue des séries temporelles, et une méthode non paramétrique.

En présence d'hétéroscédasticité, les estimateurs restent convergents mais les statistiques de tests ne sont plus distribuées selon les lois usuelles. Les tests d'autocorrélation spatiale ne sont donc plus fiables. *A contrario*, en présence d'autocorrélation spatiale, les tests d'hétéroscédasticité usuels (*White*, *Breusch-Pagan*) ne sont également plus valables. LE GALLO 2004 présente des tests joints d'hétéroscédasticité et d'autocorrélation spatiales.

La seconde forme d'hétérogénéité correspond à la variabilité spatiale des paramètres ou de la forme fonctionnelle du modèle. Lorsque l'on connait bien le territoire d'intérêt, elle est souvent traitée dans la littérature empirique en ajoutant des indicatrices de zones géographiques dans le modèle (éventuellement croisées avec chaque variable explicative), en estimant le modèle pour différentes zones ou en conduisant des tests de stabilité géographique des paramètres (dits de Chow). Lorsque le nombre de ces zones géographiques augmente, ce traitement diminue néanmoins le nombre de degrés de liberté et donc la précision des estimateurs. Des méthodes plus complexes couramment utilisées en géographie ont été développées (LE GALLO 2004). Elles restent en grande partie descriptives et exploratoires (notamment à travers des représentations graphiques), car leurs propriétés théoriques sont partiellement connues et tout particulièrement ce qui concerne les propriétés de convergence et la prise en compte des ruptures.

Il existe également des méthodes de lissage géographique où la constante (voire chaque variable explicative) est croisée avec des polynômes fonctions des coordonnées géographiques. FLACHAIRE 2005 propose un modèle linéaire partiel (et alternatif) $Y_i = X_i\beta + f(u_i, v_i) + \varepsilon_i$ où f désigne une forme fonctionnelle dépendant des coordonnées géographiques u_i et v_i (voire d'autres variables explicatives si la proximité n'est pas spatiale mais sociale ou entre produits par exemple). Il montre qu'à l'instar d'un modèle SAR, la fonction f peut s'interpréter comme une somme pondérée des variables endogènes Y. Cette analyse met ainsi en avant que corrélation et hétérogénéité spatiales sont liées.

Il existe également des méthodes de régression locale dont l'extension au contexte spatial est formalisée dans le cadre de la régression géographiquement pondérée (BRUNSDON et al. 1996). Ces méthodes sont détaillées dans le chapitre 9 : "Régression géographiquement pondérée".

Il reste néanmoins délicat de distinguer hétérogénéité et corrélation spatiales. Il n'y a pas à notre connaissance de méthode identifiant de manière distincte ces deux phénomènes. Des approches pragmatiques sont donc retenues. LE GALLO 2004 propose une application sur la criminalité aux états-Unis. À l'aide de tests d'hétéroscédasticité (robustes à la présence d'autocorrélation), elle met en avant la présence de régimes spatiaux distincts entre deux zones géographiques, Est et Ouest. Un modèle SAR est ensuite estimé, pour lequel les variables explicatives X sont croisées avec les deux régimes spatiaux et les variances sont supposées différentes entre ces deux zones. OSLAND 2010 étudie les prix de l'immobilier en Norvège à l'aide de modèles d'économétrie spatiale, de lissage semi-paramétrique et de régressions géographiques pondérées. Les différentes approches donnent des résultats complémentaires mais ne sont pas intégrées dans une modélisation unique.

6.4.4 Le risque d'erreur "écologique"

Les méthodes présentées dans ce document s'appuient sur des zonages géographiques prédéfinis (une zone d'emploi dans notre exemple). De nombreuses variables économiques ne sont disponibles que pour les divisions administratives du territoire (région, département, canton). Or ce découpage administratif ne correspond pas forcément à la réalité économique des relations entre agents. Ce phénomène géographique est connu sous l'acronyme MAUP (*Modifiable Areal Unit Problem*).

Il a plusieurs conséquences (FLOCH 2012). Avec des échelles ou des découpages différents, les résultats des modèles et les interactions entre agents ne sont pas identiques. Il faut également tenir compte de l'étendue spatiale des zones : 1 000 agents économiques n'interagissent pas de la même manière dans 1 km^2 ou dans 10 000 km^2. Lorsque des données individuelles sont disponibles (par exemple les caractéristiques d'emploi issues du recensement de la population plutôt que les taux de chômage par zone d'emploi), il est possible de faire abstraction de ce découpage administratif ou de construire le niveau géographique *a priori* le plus pertinent. Mais en général, il n'y a pas de solution pour résoudre le problème du MAUP.

De plus, les données utilisées sont souvent agrégées, au sens où elles représentent la moyenne de nos variables d'intérêt sur une zone géographique. En économétrie "classique", l'utilisation de données agrégées, connue sous le nom de *régression écologique*, entraîne des problèmes d'identification et d'hétéroscédasticité. Anselin (2002) donne l'exemple d'un modèle où les décisions d'un individu i, y_{ik}, s'expliquent par ses caractéristiques x_{ik} mais également par les caractéristiques du groupe k auquel il appartient $\bar{x}_k = \sum_i x_{ik}/n_k$. Le modèle s'écrit $y_{ik} = \alpha + \beta \cdot x_{ik} + \gamma \cdot \bar{x}_k + \varepsilon_{ik}$ où β représente l'effet individuel et γ l'effet de contexte. Si on ne dispose que de données par groupe (par exemple les résultats moyens d'une classe à un examen et non les résultats individuels), le modèle estimé devient $\bar{y}_k = \alpha + (\beta + \gamma) \cdot \bar{x}_k + \bar{\varepsilon}_k$. Il n'est alors plus possible d'identifier séparément les paramètres β et γ. Le modèle est hétéroscédastique car $\mathbb{V}(\bar{\varepsilon}_k) = \sigma^2/n_k$ dans le cas de perturbations initiales i.i.d. de variance σ^2.

Le problème est encore plus complexe dans le cas de modèles spatiaux. Il n'est en effet pas possible d'agréger une matrice de voisinage W définie au niveau individuel. Avec des données individuelles, un individu i du groupe k peut avoir des voisins parmi le groupe k mais également parmi un autre groupe k'. Si on considère désormais une matrice de voisinage agrégée au niveau groupe, les relations intra-groupe ne seront plus prises en compte (la diagonale est nulle par hypothèse). De plus, il peut y avoir de nombreux individus du groupe k voisins d'individus du groupe k' mais très peu voisins d'un autre groupe k''. Avec une matrice de contiguïté agrégée au niveau groupe, la force des relations individuelles ne sera plus prise en compte (chaque voisin a le même poids). Au-delà des problèmes d'identification d'une *régression écologique*, un modèle SAR défini au niveau individuel ne peut pas être agrégé pour correspondre à un modèle SAR défini à un niveau supérieur. Il n'y a pas de relations simples entre les paramètres.

Pour bien comprendre cette question, prenons l'exemple du marché immobilier. On observe des villes dont les prix sont très élevés au centre et diminuent ensuite progressivement. Il existe également des niveaux de prix très différents entre les villes. Si on ne considère que des prix moyens par centre urbain (regroupant des villes proches), la disparité des prix au sein des villes sera cachée. Ces emboîtements d'échelles peuvent engendrer des résultats à première vue paradoxaux.

En pratique, cela signifie que l'interprétation des résultats n'est valable que pour le découpage géographique choisi. Si on étudie des relations économiques à un niveau agrégé avec un modèle spatial, on ne peut rien dire des relations individuelles entre agents. Pour tenir compte de cette imbrication des zones géographiques (régions, départements, cantons, individus) et rendre les analyses cohérentes entre elles, une solution est alors de mener des analyses multi-niveaux (GIVORD et al. 2016). Dans le cas d'études macro-économiques telles que la croissance régionale, ce problème est moins présent. Le niveau agrégé est en effet le niveau pertinent.

6.5 Mise en pratique sous R

Dans cette partie, nous détaillons la mise en pratique d'une étude d'économétrie spatiale, en modélisant le taux de chômage localisé (par zone d'emploi, hors Corse) à l'aide de caractéristiques structurelles relatives aux caractéristiques de la population active (proportion des personnes peu

diplômées et des personnes de moins de 30 ans dans la population active), de la structure économique (proportion des emplois dans le secteur industriel et de l'emploi public) et du marché du travail (taux d'activité). L'objectif de cette partie n'est pas de détailler les résultats d'une étude économique mais d'illustrer les techniques mises en œuvre : la définition d'une matrice de voisinage qui décrit les relations de proximité entre territoires, les tests de corrélation spatiale et de spécification, l'estimation, et l'interprétation de modèles d'économétrie spatiale. D'autres variables peuvent bien sûr expliquer les taux de chômage locaux (BLANC et al. 2008 ; LOTTMANN 2013). Les variables économiques sont supposées structurelles et peu variables à court terme. Pour limiter les problèmes d'endogénéité, le taux de chômage est calculé sur l'année 2013 et les variables explicatives correspondent aux millésimes 2011 de la source CLAP (Connaissance locale de l'Appareil Productif) et du RP (Recensement de la Population). Une interprétation causale reste néanmoins impossible. De nombreuses variables ont en effet été omises de l'analyse, comme par exemple l'offre d'emploi. Les variables explicatives prises en compte peuvent ainsi intégrer l'effet de ces variables omises et non leur seul effet propre. Enfin, le décalage temporel entre les variables explicatives et le taux de chômage ne supprime pas complètement le caractère simultané des phénomènes (par exemple entre le taux d'activité et le taux de chômage), structurellement stables à court terme.

Les exemples et les codes sont présentés à l'aide de R, logiciel le plus complet pour l'estimation de modèles d'économétrie spatiale. Nous listons ci-dessous quelques packages utiles dans R :

— *sp* et *rgdal* pour l'importation et la définition des objets spatiaux, *maptools* pour la définition de cartes ;
— des fonctions similaires à celles de SIG (Système d'Information Géographique) du type calcul de distance ou des méthodes de géostatistique : *fields*, *raster* et *gdistance* ;
— l'économétrie spatiale : *spdep* (spatial dependancies) pour l'ensemble des modèles classiques, et *spgwr* pour la régression géographique pondérée.

6.5.1 Cartographie et tests

Après avoir importé les données et défini une matrice de voisinage grâce aux méthodes présentées en section 6.2, on peut cartographier les données et effectuer une première analyse de l'autocorrélation spatiale.

La figure 6.4 représente les taux de chômage par zone d'emploi en 2013. On constate des zones polarisées, ce qui pourrait être le signe d'une hétérogénéité spatiale. Le Nord de la France et le Languedoc-Roussillon présentent ainsi des taux de chômage plus élevés, les zones frontalières de la Suisse plus faibles. Les zones contiguës de ces régions ont des taux de chômage proches également, ce qui est caractéristique d'une autocorrélation spatiale. Pour les variables explicatives, on constate notamment une polarisation forte du pourcentage d'emploi industriel. Les taux d'activité présentent une structuration spatiale proche du taux de chômage.

La table 6.1 décrit la distribution des variables. Le taux de chômage moyen est de 10%, pour un taux d'activité de 73%. Il y a 22% d'actifs peu diplômés et de jeunes actifs de moins de 30 ans. Hormis pour le pourcentage d'emploi industriel et d'emploi public, les écarts interquartiles sont faibles, inférieurs à 5%. Le pourcentage d'emploi industriel apparaît comme la variable la plus polarisée.

Tests d'autocorrélation spatiale et représentations graphiques avancées

La p-value quasiment nulle du test de Moran indique que l'hypothèse nulle d'absence d'autocorrélation spatiale doit être rejetée (voir chapitre 3 : "Indices d'autocorrélation spatiale"). Le résultat est robuste au choix de la matrice de voisinage.

L'autocorrélation des données brutes peut être illustrée graphiquement à l'aide du graphique de Moran. Il met en relation la valeur observée en un point et celle qui est observée dans le voisinage déterminé par la matrice de poids.

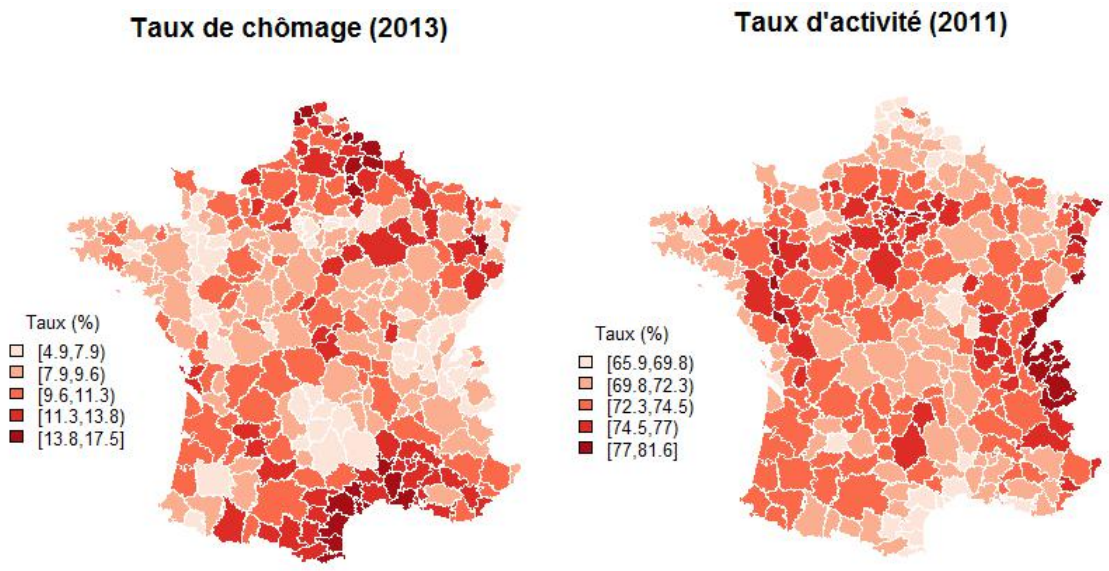


FIGURE 6.4 – Distribution du taux de chômage et d'activité, par zone d'emploi

	N	Moyenne	Écart-Type	Min	Q25	Médiane	Q75	Max
Taux de chômage (en %)	297	10.0	2.4	4.9	8.3	9.6	11.4	17.5
Taux d'activité (en %)	297	72.8	2.6	65.9	71.3	72.8	74.2	81.6
% Actifs Peu Diplômés	297	22.1	3.6	13.0	19.5	22.2	24.8	32.2
% Jeunes Actifs 15-30 ans	297	21.8	2.0	16.7	20.4	21.8	23.2	27.7
% Emploi Industriel	297	19.7	8.8	3.7	13.3	18.2	24.8	52.0
% Emploi Public	297	33.5	6.2	15.0	29.5	33.2	36.9	51.0

TABLE 6.1 – Descriptif de l'échantillon
Note : La zone géographique est la zone d'emploi. Les statistiques ne sont pas pondérées.

La figure 6.5 est cohérente avec les résultats du test de Moran. Une relation linéaire apparaît entre le taux de chômage d'une zone et celui de son voisinage. Une carte peut être associée permettant de situer les zones d'emploi en fonction de leurs caractéristiques (HH signifie un taux de chômage élevé dans un environnement élevé, HB un taux élevé dans un environnement plus bas). Elle permet de constater que cette relation n'est pas homogène sur le territoire. Le Nord et le Sud présentent des taux de chômage élevés. Une France du "milieu" présente au contraire des taux de chômage moindres.

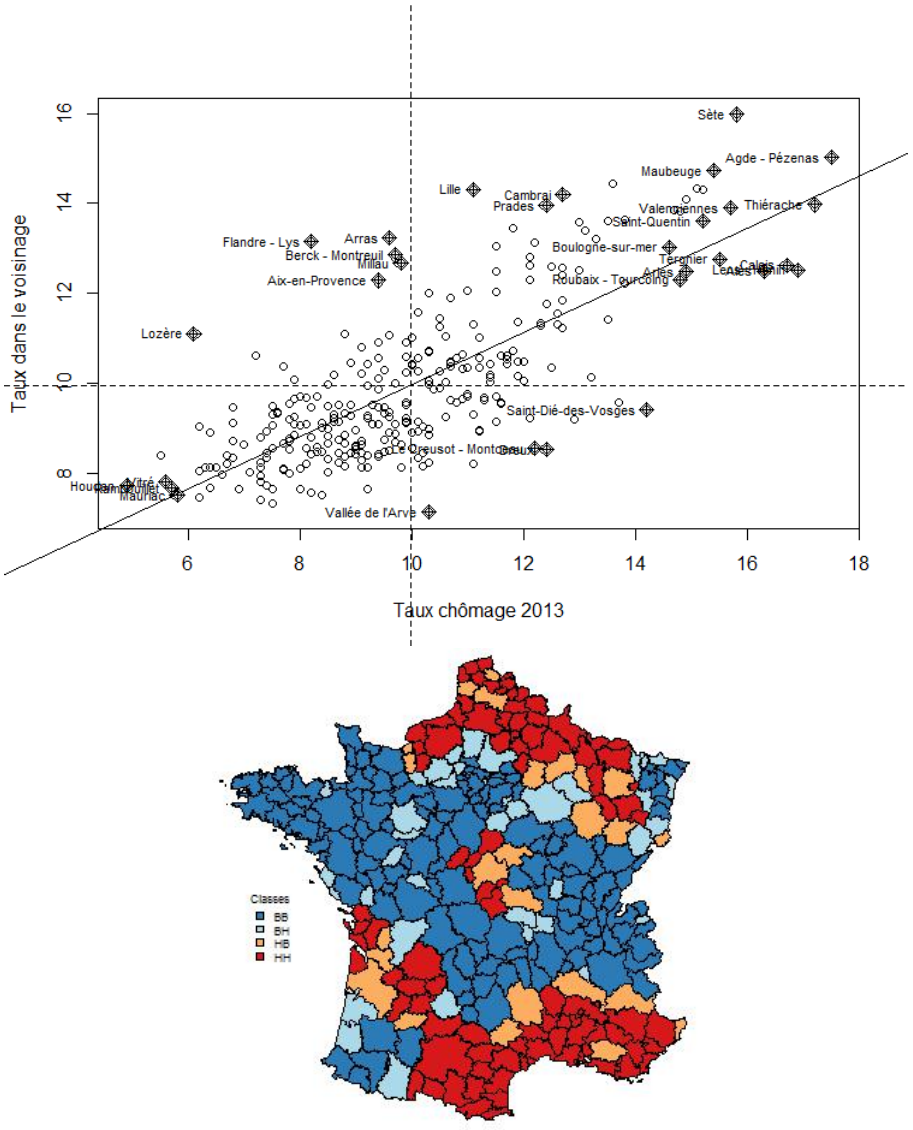


FIGURE 6.5 – Graphique de Moran du taux de chômage et carte associée

6.5.2 Estimation et choix de modèles

L'analyse descriptive a permis de constater que l'espace n'était pas neutre pour caractériser les taux de chômage locaux. Il n'est néanmoins pas certain qu'un modèle économétrique tenant compte de l'espace soit nécessaire. Le nuage de points des taux de chômage et d'activité montrent une forte relation linéaire des deux variables. Le taux de chômage et d'activité sont tous les deux corrélés spatialement. Le taux de chômage pourrait donc être relié au taux d'activité, sans autre forme de

corrélation spatiale que celle présente dans les deux variables. Première chose, on commence donc par estimer un modèle linéaire non spatial à l'aide des MCO. Un test de Moran adapté sur les résidus confirme la présence résiduelle d'autocorrélation spatiale (potentiellement associée à de l'hétérogénéité spatiale), quelle que soit la matrice de voisinage.

Pour déterminer la forme de corrélation spatiale (endogène, exogène ou inobservée), la démarche est pragmatique. L'approche de ELHORST 2010 conduirait à retenir le modèle SDM. Seuls les modèles MCO et SDM seraient alors estimés. Dans un but pédagogique, l'ensemble des modèles spatiaux est néanmoins estimé, pour 6 matrices de voisinage : contiguïté, plus proches voisins (2, 5 ou 10), distance inverse, et proportionnelle aux trajets domicile-travail (dite matrice endogène). Les régressions s'estiment à l'aide du package *spdep*. Le coût computationnel à estimer ces modèles est par ailleurs faible.

```
### Modèle estimé
modele <- txcho_2013 ~ tx_act+part_act_peudip+part_act_1530+part_emp_ind+
    part_emp_pub
### Matrice de voisinage
matrice <- dist.w

### Modèle MCO
ze.lm <- lm(modele, data=donnees_ze)
summary(ze.lm)

### Test de Moran adapté sur les résidus
lm.morantest(ze.lm,matrice)

### Test LM-Error et LM-Lag
lm.LMtests(ze.lm,matrice,test="LMerr")
lm.LMtests(ze.lm,matrice,test="LMlag")
lm.LMtests(ze.lm,matrice,test="RLMerr")
lm.LMtests(ze.lm,matrice,test="RLMlag")

### Modèle SEM
ze.sem<-errorsarlm(modele, data=donnees_ze, matrice)
summary(ze.sem)
### Test d'Hausman
Hausman.test(ze.sem)

### Modèle SAR
ze.sar<-lagsarlm(modele, data=donnees_ze, matrice)
summary(ze.sar)

### Modèle SDM
ze.sardm<-lagsarlm(modele, data=donnees_ze, matrice, type="mixed")
summary(ze.sardm)
### Test de l'hypothèse de facteur commun
# ze.sardm : Modèle non contraint
# ze.sem : Modèle contraint
FC.test<-LR.sarlm(ze.sardm,ze.sem)
print(FC.test)
```

On ne présente ici que les résultats associés à la matrice de distances inverse, car c'est celle qui présente le caractère explicatif le plus fort (AIC les plus faibles) et dont l'interprétation économique est la plus intuitive. Les zones d'emploi n'ayant pas la même taille, la contiguïté ou les plus proches voisins peuvent engendrer des effets inattendus. La matrice endogène peut par construction provoquer un biais des estimateurs. Les résultats sur le choix de modèle restent néanmoins cohérents, quelle que soit la matrice de voisinage retenue.

Nous nous attendons ici à une relation négative entre taux de chômage et taux d'activité, mais positive pour le pourcentage d'actifs peu diplômés et de jeunes actifs. Le "halo" du chômage est moins présent dans les zones dynamiques en termes d'emploi. Les personnes les moins diplômées et les jeunes sont réputés plus touchés par le chômage. Les zones de fort emploi industriel sont *a priori* plus affectées par le chômage (réaction de l'emploi à la conjoncture et fermeture d'usines). Au contraire, les emplois publics étant plus stables, le pourcentage d'emploi public devrait être négativement corrélé avec le taux de chômage. Rappelons ici que ce modèle se veut illustratif des techniques d'économétrie spatiale, aucune conclusion économique ne peut en être tirée.

Concernant le choix du modèle, on peut retenir les points suivants de la table 6.2.

— L'approche séquentielle d'Elhorst (présentée en 6.3.2) conduirait à retenir un modèle SDM (colonne 4). Il présente l'AIC le plus faible (960). L'ensemble des tests d'autocorrélation spatiale menés à partir des résidus du modèle MCO sont rejetés (colonne 1). De même, l'hypothèse de facteur commun du modèle SDM est rejetée (p-value de 0.004). Plusieurs effets d'interaction exogène sont significativement non nuls (le pourcentage d'actifs non diplômés au seuil de 1 %). Enfin, pour le modèle à interactions exogènes (SLX, colonne 6), on ne rejette pas l'hypothèse d'absence d'autocorrélation résiduelle sous l'hypothèse de corrélation endogène (test robuste LM-Error, p-value de 0.787).

— Le choix d'un modèle SAR (colonne 3) serait ici déconseillé. Un test montre qu'une autocorrélation spatiale résiduelle reste présente (p-value (test LM residual auto.) de 0.003). Les conséquences sont importantes sur l'interprétation des résultats. La variable "pourcentage d'emploi industriel" reste significative à 1 % (quelle que soit la matrice de voisinage), alors que le signe négatif peut paraître contre-intuitif.

— Le modèle de Manski (colonne 8) fournit des résultats divergents selon la matrice de voisinage (non présentés ici), certainement par manque d'identifiabilité de ce modèle. De même, le modèle SAC (corrélation endogène et résiduelle, colonne 5) estime une corrélation endogène faible et non significative en comparaison de l'autocorrélation résiduelle. Ce résultat est difficile à interpréter et peut provenir d'une mauvaise spécification du modèle (Le Gallo 2002).

Enfin, pour des raisons de parcimonie, le choix d'un modèle SEM (table 6.2, colonne 2) voire SDEM (colonne 7) pourrait être envisagé, après avoir vérifié la cohérence des résultats avec ceux du modèle SDM. L'interprétation de ce modèle SEM est en effet plus aisée mais se limite aux effets directs. Le critère AIC (967) est proche du modèle SDM, et pour des matrices de poids des 5 ou 10 plus proches voisins (table 6.3, colonnes 4 et 5), l'hypothèse de facteur commun n'est pas rejetée à 1 %. La divergence de résultats entre les modèles MCO et SEM pourrait amener à conclure que la spécification du modèle SEM n'est pas juste, *i.e.* qu'elle souffre d'un biais de variable omise. Un test d'Hausman (LeSage et Pace 2009 p.61-63) entre les modèles MCO et SEM repose sur l'hypothèse nulle de validité des deux modèles, le modèle SEM étant plus efficace. On constate alors que cette hypothèse n'est pas rejetée au seuil de 1 %, hormis pour la matrice de poids des 2 plus proches voisins (table 6.3).

Les divergences de résultats (pour différentes matrices de voisinage) sont analysées pour les modèles SEM et SDM. Le modèle SEM peut s'interpréter comme le modèle MCO. L'effet marginal correspond bien aux paramètres du modèle. Cette comparaison est cohérente avec un biais du modèle MCO. Pour le taux d'activité, l'effet est surévalué de 0.09 à 0.12 point par rapport au

	(1) MCO	(2) SEM	(3) SAR	(4) SDM	(5) SAC	(6) SLX	(7) SDEM	(8) Manski
Taux d'activité	-0.622***	-0.498***	-0.437***	-0.472***	-0.499***	-0.470***	-0.486***	-0.473***
	(0.039)	(0.041)	(0.038)	(0.042)	(0.041)	(0.050)	(0.041)	(0.042)
% Actifs Peu Diplômés	0.186***	0.184***	0.138***	0.182***	0.179***	0.179***	0.181***	0.183***
	(0.026)	(0.027)	(0.022)	(0.027)	(0.026)	(0.033)	(0.027)	(0.028)
% Jeunes Actifs 15-30 ans	0.138***	0.196***	0.087**	0.209***	0.180***	0.205***	0.197***	0.211***
	(0.043)	(0.045)	(0.037)	(0.046)	(0.045)	(0.055)	(0.045)	(0.047)
% Emploi Industriel	-0.062***	-0.018	-0.036***	-0.015	-0.021*	-0.022	-0.024**	-0.014
	(0.012)	(0.012)	(0.010)	(0.012)	(0.012)	(0.014)	(0.012)	(0.012)
% Emploi Public	-0.068***	-0.044***	-0.063***	-0.042**	-0.048***	-0.044**	-0.049***	-0.041**
	(0.019)	(0.016)	(0.016)	(0.016)	(0.017)	(0.019)	(0.017)	(0.016)
$\hat{\rho}$			0.519***	0.629***	0.205*			0.689***
			(0.049)	(0.064)	(0.109)			(0.120)
$\hat{\lambda}$		0.747***			0.616***		0.651***	-0.137
		(0.051)			(0.096)		(0.063)	(0.257)
$\hat{\theta}$, Taux d'activité				0.157*		-0.300***	-0.277***	0.205*
				(0.083)		(0.082)	(0.105)	(0.111)
$\hat{\theta}$, % Actifs Peu Diplômés				-0.135***		-0.027	-0.021	-0.145***
				(0.045)		(0.052)	(0.066)	(0.046)
$\hat{\theta}$, % Jeunes Actifs 15-30 ans				-0.140 *		-0.041	-0.003	-0.153**
				(0.072)		(0.085)	(0.115)	(0.072)
$\hat{\theta}$, % Emploi Industriel				-0.044**		-0.118***	-0.073**	-0.038*
				(0.020)		(0.023)	(0.029)	(0.023)
$\hat{\theta}$, % Emploi Public				-0.024		-0.084*	-0.070	-0.018
				(0.037)		(0.043)	(0.052)	(0.037)
Constante	51.653***	39.729***	34.470***	27.456***	38.427***	66.077***	63.650***	23.530***
	(3.635)	(3.685)	(3.407)	(6.766)	(3.901)	(6.514)	(10.213)	(9.065)
Observations	297	297	297	297	297	297	297	297
AIC	1072	967	980	960	967	1029	964	962
R^2 Ajusté	0.624					0.679		
Test Moran	0.000					0.000		
Test LM-Error	0.000					0.000		
Test LM-Lag	0.000					0.000		
Test Robuste LM-Error	0.000					0.787		
Test Robuste LM-Lag	0.000					0.001		
Test Facteur Commun				0.004				
Test LM residual auto.			0.003	0.572				

TABLE 6.2 – Déterminants du taux de chômage par zone d'emploi, à partir d'une matrice inverse de la distance

Note : L'ensemble des modèles est estimé avec une matrice inverse de la distance (avec un seuil à 100 km). Les écarts-types sont indiqués entre parenthèses. Pour les tests, la p-value est indiquée. Significativité : * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

modèle SEM. Pour le pourcentage d'emploi industriel, le modèle MCO conclut à un effet négatif significatif alors qu'il est jugé nul avec le modèle SEM dans le cas d'une matrice inverse de la distance ou plus faible avec les autres matrices. L'effet du taux d'activité pourrait être surévalué avec une matrice de contiguïté ou un nombre faible de plus proches voisins. L'effet du pourcentage de jeunes actifs semble sous-évalué avec une matrice endogène. Pour le modèle SDM (table 6.7 en annexe de ce chapitre), une interprétation directe n'est pas possible car les effets doivent tenir compte des effets d'interaction endogène. On constate des effets d'interactions exogènes variables selon la matrice de voisinage.

Les résultats pour le modèle SEM ne sont pas toujours robustes au choix de la matrice de voisinage, le "pourcentage d'emploi industriel" pouvant se révéler ou non significatif. Il n'y a pas de choix évident de matrice de voisinage, qui amènerait à privilégier les résultats obtenus avec une matrice inverse de la distance par exemple. Le choix ne doit bien sûr en aucun cas être dicté par un argument de significativité des résultats, mais reposer sur une analyse associée à la question économique.

	(1) MCO	(2) SEM Contiguïté	(3) SEM 2 Voisins	(4) SEM 5 Voisins	(5) SEM 10 Voisins	(6) SEM Distance	(7) SEM Endogène
Taux d'activité	-0.622***	-0.518***	-0.517***	-0.530***	-0.507***	-0.498***	-0.515***
	(0.039)	(0.040)	(0.040)	(0.040)	(0.040)	(0.041)	(0.041)
% Actifs Peu Diplômés	0.186***	0.188***	0.204***	0.185***	0.181***	0.184***	0.184***
	(0.026)	(0.026)	(0.026)	(0.026)	(0.026)	(0.027)	(0.026)
% Jeunes Actifs 15-30 ans	0.138***	0.179***	0.195***	0.201***	0.198***	0.196***	0.139***
	(0.043)	(0.045)	(0.044)	(0.045)	(0.046)	(0.045)	(0.044)
% Emploi Industriel	-0.062***	-0.023*	-0.027**	-0.023*	-0.024**	-0.018	-0.026**
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
% Emploi Public	-0.068***	-0.042**	-0.039**	-0.047***	-0.048***	-0.044***	-0.050***
	(0.019)	(0.017)	(0.017)	(0.017)	(0.017)	(0.016)	(0.016)
$\hat{\lambda}$		0.687***	0.506***	0.681***	0.763***	0.747***	0.700***
		(0.050)	(0.047)	(0.051)	(0.053)	(0.051)	(0.044)
Constante	51.653***	41.535***	40.672***	42.166***	40.685***	39.729***	42.414***
	(3.635)	(3.681)	(3.643)	(3.639)	(3.644)	(3.685)	(3.745)
Observations	297	297	297	297	297	297	297
AIC	1072	977	996	972	973	967	995
Test Hausman		0.030	0.000	0.042	0.114	0.029	0.115
Test Facteur Commun		0.002	0.001	0.040	0.035	0.004	0.000

TABLE 6.3 – Modèle SEM, pour différentes matrices de voisinage
Note : Le modèle SEM est estimé avec 6 matrices de voisinage différentes. Les écarts-types sont indiqués entre parenthèses. Significativité : * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

6.5.3 Interprétation des résultats

Pour le modèle SDM, afin de permettre une interprétation au regard du modèle MCO et SEM, on calcule les effets directs et indirects tels que décrits en section 6.4 (tables 6.4 et 6.5). Les intervalles de confiance empiriques sont obtenus à l'aide de 1 000 simulations à partir de la distribution empirique. Pour les effets directs, on retrouve l'interprétation du modèle SEM. Pour les effets indirects, seul le pourcentage d'emploi industriel a un effet négatif significatif. Ces effets indirects ont en effet une variabilité plus grande, qui ne permet pas de conclure sur les effets éventuels. Le modèle SDM met en avant le rôle particulier du pourcentage d'emploi industriel, qui seul aurait un effet indirect (négatif) associé à un effet direct (négatif) faible ou nul selon la matrice de voisinage retenue. La compréhension économique d'un tel résultat demeure délicate. Le modèle SDM peut amener à interpréter de manière fallacieuse la corrélation endogène, qui n'a pas ici une interprétation économique claire. Au vu de ces résultats, le modèle SEM pourrait ainsi être privilégié par principe de parcimonie.

```
### Estimation des effets directs et indirects du modèle SDM
```

```
impactssdm<-impacts(ze.sardm, listw=matrice, R=1000)
summary(impactssdm)
```

	(1) MCO	(2) SDM Contiguïté	(3) SDM 2 Voisins	(4) SDM 5 Voisins	(5) SDM 10 Voisins	(6) SDM Distance	(7) SDM Endogène
Taux d'activité	-0.622	-0.509	-0.510	-0.529	-0.505	-0.490	-0.508
	[-0.700,-0.545]	[-0.588,-0.435]	[-0.589,-0.434]	[-0.611,-0.451]	[-0.583,-0.422]	[-0.574,-0.409]	[-0.588,-0.429]
% Actifs Peu Diplômés	0.186	0.178	0.208	0.183	0.177	0.180	0.178
	[0.136,0.237]	[0.122,0.232]	[0.154,0.261]	[0.132,0.235]	[0.125,0.230]	[0.122,0.230]	[0.129,0.232]
% Jeunes Actifs 15-30 ans	0.138	0.194	0.223	0.213	0.212	0.207	0.184
	[0.054,0.223]	[0.102,0.288]	[0.135,0.312]	[0.123,0.309]	[0.119,0.306]	[0.119,0.299]	[0.092,0.279]
% Emploi Industriel	-0.062	-0.026	-0.032	-0.027	-0.027	-0.022	-0.033
	[-0.087,-0.038]	[-0.048,-0.003]	[-0.053,-0.008]	[-0.051,-0.005]	[-0.050,-0.005]	[-0.045,0.001]	[-0.055,-0.011]
% Emploi Public	-0.068	-0.045	-0.048	-0.052	-0.051	-0.049	-0.052
	[-0.106,-0.030]	[-0.078,-0.010]	[-0.081,-0.011]	[-0.084,-0.017]	[-0.083,-0.018]	[-0.081,-0.014]	[-0.084,-0.019]

TABLE 6.4 – Impacts directs du modèle SDM, pour différentes matrices de voisinage
Note : Le modèle SDM est estimé avec 6 matrices de voisinage différentes. Les intervalles de confiance empiriques (quantiles à 2,5 % et 97,5 % de 1000 simulations MCMC) sont indiqués entre crochets.

	(1) SDM Contiguïté	(2) SDM 2 Voisins	(3) SDM 5 Voisins	(4) SDM 10 Voisins	(5) SDM Distance	(6) SDM Endogène
Taux d'activité	-0.323	-0.200	-0.241	-0.306	-0.357	-0.351
	[-0.587,-0.091]	[-0.337,-0.068]	[-0.488,0.007]	[-0.700,0.030]	[-0.658,-0.073]	[-0.638,-0.107]
% Actifs Peu Diplômés	-0.015	-0.059	-0.032	-0.050	-0.053	-0.079
	[-0.161,0.142]	[-0.146,0.032]	[-0.205,0.124]	[-0.291,0.158]	[-0.254,0.137]	[-0.251,0.085]
% Jeunes Actifs 15-30 ans	-0.016	-0.079	-0.082	0.016	-0.023	0.047
	[-0.321,0.249]	[-0.214,0.058]	[-0.334,0.174]	[-0.321,0.390]	[-0.352,0.301]	[-0.230,0.332]
% Emploi Industriel	-0.130	-0.064	-0.100	-0.135	-0.136	-0.111
	[-0.208,-0.055]	[-0.105,-0.022]	[-0.170,-0.030]	[-0.244,-0.041]	[-0.229,-0.059]	[-0.187,-0.043]
% Emploi Public	-0.120	-0.078	-0.113	-0.098	-0.130	-0.037
	[-0.274,0.017]	[-0.140,-0.011]	[-0.257,0.031]	[-0.345,0.132]	[-0.335,0.046]	[-0.186,0.106]

TABLE 6.5 – Impacts indirects du modèle SDM, pour différentes matrices de voisinage
Note : Le modèle SDM est estimé avec 6 matrices de voisinage différentes. Les intervalles de confiance empiriques (quantiles à 2,5 % et 97,5 % de 1000 simulations MCMC) sont indiqués entre crochets.

6.5.4 Autres modélisations spatiales

L'analyse descriptive a mis en avant une hétérogénéité spatiale possible du modèle. Il serait possible d'intégrer et de tester la présence de ce phénomène, soit en autorisant le modèle à être hétéroscédastique (*via* le package *sphet*, PIRAS et al. 2010), soit en modélisant une variabilité spatiale des paramètres ou de la forme fonctionnelle du modèle. Cette seconde forme d'hétérogénéité est obtenue en incluant des indicatrices de zones géographiques dans le modèle, à l'aide d'un modèle de lissage géographique (*via* le package *McSpatial*, qui inclut des modèles spatiaux semi-paramétriques ou par splines) ou en conduisant une analyse géographique pondérée.

La mise en œuvre pratique d'une régression géographiquement pondérée est détaillée dans le chapitre 9 : "Régression géographiquement pondérée". Nous présentons ici les résultats de l'estimation géographiquement pondérée du modèle linéaire reliant le taux de chômage et les caractéristiques structurelles présenté plus haut.

La table 6.6 fournit les valeurs minimales, maximales et les quartiles des coefficients obtenus. On peut ainsi apprécier la variabilité des coefficients, et comparer ces résultats avec ceux des MCO. L'utilisation de la régression géographique conduit à des coefficients qui ne sont pas toujours de même signe. Cela peut conduire à s'interroger sur le bien-fondé de la spécification. Les coefficients peuvent varier de façon sensible, notamment pour les actifs de 15 à 30 ans, le coefficient médian s'écartant très sensiblement de celui des MCO.

	(1) MCO	(2) Minimum	(3) P1	(4) Médiane	(5) P3	(6) Maximum
Taux d'activité	-0.622	-1.492	-0.653	-0.508	-0.379	-0.133
% Actifs Peu Diplômés	0.186	-0.116	0.081	0.188	0.250	0.607
% Jeunes Actifs 15-30 ans	0.138	-0.753	-0.040	0.183	0.340	0.875
% Emploi Industriel	-0.062	-0.233	-0.066	-0.029	0.006	0.184
% Emploi Public	-0.068	-0.318	-0.098	-0.048	-0.002	0.218
Constante	51.650	-7.485	29.940	40.440	52.310	130.500

TABLE 6.6 – Résultats de la régression géographique pondérée

On récupère une table contenant pour chacun des points d'estimation (ici les centroïdes des zones d'emploi) la valeur des coefficients, la valeur prédite par le modèle, les résidus et la valeur locale du R^2. Cela permet notamment de cartographier les variations locales des paramètres. Cette dimension cartographique est importante pour apprécier les tendances spatiales. On peut également vérifier si les résidus restent autocorrélés spatialement, à l'aide de cartes et de tests de Moran adaptés. Il n'y a pas une structure spatiale marquée des résidus dans le cas présent. La distribution des paramètres spatiaux pour la part de l'emploi industriel et de l'emploi public (figure 6.6) met en avant des particularités régionales, qui peuvent permettre de comprendre des résultats surprenants, par exemple la relation nulle (ou négative) entre emploi industriel et taux de chômage. Cette relation négative est présente principalement dans la partie Sud de la France (ainsi que quelques zones du Nord), alors que des zones du Centre et de l'Est, régions ayant subi de fortes restructurations industrielles, présentent une corrélation positive entre taux de chômage et part de l'emploi industriel. Concernant l'emploi public, on constate une relation négative avec le taux de chômage pour une partie du Sud de la France et du Nord, alors que la relation est positive en Bretagne par exemple. Notre modèle inclut un nombre limité de variables, l'effet de certaines particularités régionales (restructurations industrielles, caractéristiques de l'offre d'emploi, etc.) pourrait ainsi être capté à tort par nos variables explicatives, biais classique d'endogénéité. Il est également possible que les comportements soient hétérogènes entre zones d'emploi. Dans tous les cas, cette analyse devrait nous amener à modifier notre modèle, par l'inclusion d'autres variables ou de paramètres de corrélation spatiale par zones géographiques. Nous limitons ici notre analyse, en rappelant que les résultats présentés ne visent qu'à illustrer la démarche du choix et de l'estimation d'un modèle spatial. Prendre en compte à la fois l'hétérogénéité et la corrélation spatiales demeure délicat.

Nous avons effectué les tests permettant de vérifier la non-stationnarité, et donc d'apprécier si la régression géographique pondérée est préférable au modèle linaire estimé par les MCO (BRUNSDON et al. 2002 ; LEUNG et al. 2000). La stationnarité est rejetée ici quel que soit le test, au niveau global et pour chaque variable explicative (résultats non présentés ici).

La régression géographique pondérée est considérée comme une bonne méthode exploratoire,

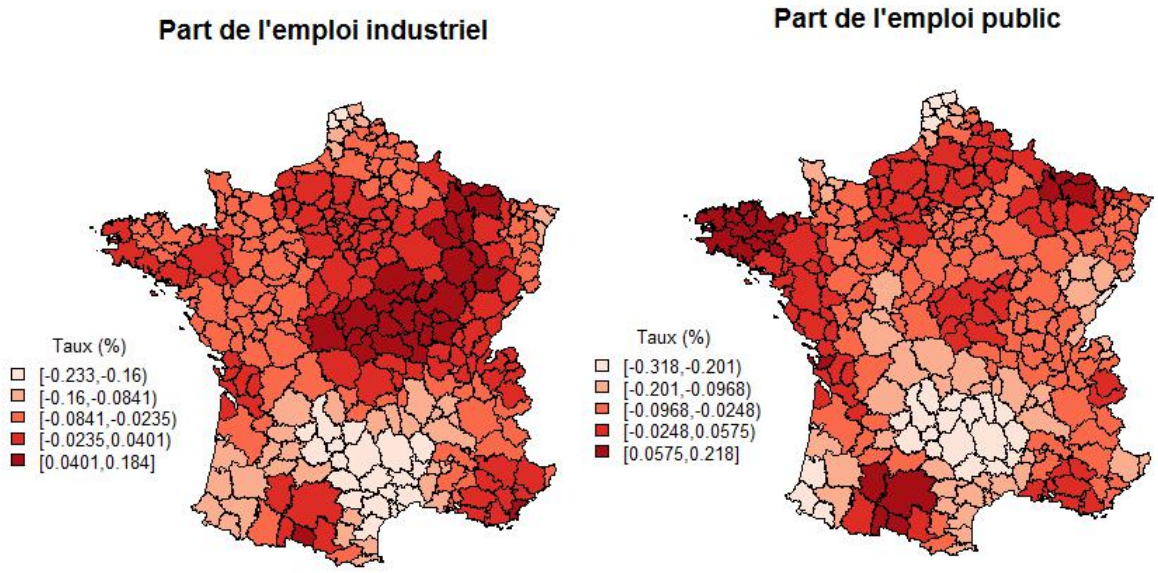


FIGURE 6.6 – Distribution des paramètres locaux

permettant notamment de visualiser des phénomènes de non-stationnarité. Mais elle a fait aussi l'objet d'un certain nombre de critiques. WHEELER et al. 2009 soulignent que les résultats ne sont pas robustes à une forte corrélation entre variables explicatives ou à la présence conjointe d'autocorrélation spatiale. De plus, comme dans toutes les méthodes statistiques non paramétriques, la distance introduite (*i.e.* le choix de la fenêtre) n'est pas neutre. Une grande distance, introduisant de nombreux points, va conduire à des coefficients variant peu localement. À l'inverse une faible distance introduira beaucoup de variabilité. Le choix opéré peut avoir des conséquences sur les tests appréciant le choix de la régression géographique pondérée par rapport aux MCO. Le package *GWmodel* (BRUNSDON et al. 2015) tente de répondre à ces critiques.

Conclusion

Les modèles d'économétrie spatiale définissent un cadre cohérent (et paramétrique) pour modéliser tout type d'interactions entre agents économiques : zones géographiques mais également produits, entreprises ou individus. Ils reposent sur une définition *a priori* de relations de voisinage. Les principales critiques qui leur sont adressées sont leur manque de robustesse quant au choix de la matrice de voisinage et leur manque d'identification du processus générateur des données. Ces critiques nous semblent néanmoins exagérées. Comme pour tout travail empirique, des choix toujours discutables de spécification sont nécessaires. La force de ces modèles est de mettre en avant si un problème "spatial" se pose et sous quelle forme. *A contrario*, estimer un modèle d'économétrie spatiale dès qu'on dispose de données "spatiales" n'est pas toujours nécessaire. Le raffinement méthodologique doit être mis en regard de la question économique et de la complexité de ces nouveaux modèles, en particulier en termes d'interprétation.

Le choix de modéliser la corrélation ou l'hétérogénéité spatiale, voire les deux simultanément, est délicat. Dans notre exemple, prendre en compte la corrélation spatiale pour modéliser le taux de chômage localisé apparaît nécessaire d'après les tests statistiques. Cela corrige certaines interprétations erronées issues du modèle linéaire classique. Il conviendrait ici de privilégier un modèle spatial de Durbin (SDM), voire aux erreurs spatialement autocorrélées (modèle SEM). Mais l'analyse de l'hétérogénéité spatiale à partir de régressions géographiques pondérées souligne également que la spécification devrait être améliorée, certains résultats surprenants pouvant provenir

d'un biais de variables omises et d'une mauvaise prise en compte de l'hétérogénéité spatiale des marchés du travail. Cette incertitude sur le choix du modèle doit amener à rester prudent quant à l'interprétation des effets directs et indirects du modèle SDM. De plus, ce n'est pas parce que le modèle est plus compliqué qu'il règle le problème de l'endogénéité des variables explicatives ou du sens de la causalité entre les variables du modèle. Aucune interprétation causale n'est ici possible.

Les enjeux théoriques de ces méthodes, et en particulier les liens entre corrélation et hétérogénéité spatiales, ne sont pas complètement maîtrisés. Les modèles d'économétrie spatiale permettent une prise en compte de l'espace ou des relations entre agents, préférable bien souvent à ne rien faire. La régression géographique pondérée et le lissage géographique permettent en complément des approches descriptives, de définir de grands ensembles régionaux homogènes et des analyses complémentaires à des tests de rupture régionale. Néanmoins, estimer ces modèles suppose de disposer de données exhaustives. Dans le cas général, ils ne sont donc pas adaptés aux données d'enquêtes.

Annexes

Annexe 1 : Codes R complémentaires

Création d'une matrice de voisinage endogène, basée sur les déplacements domicile-travail

```
## Lecture du fichier SAS, des flux domicile-travail
library ("sas7bdat")
flux<-read.sas7bdat("flux.sas7bdat")
## Numérotation des zones
zeo<-unique(flux[,1])
zed<-unique(flux[,1])
lig<-c(rep(1:297))
col<-c(rep(1:297))
dzeo<-data.frame(zeo,lig)
dzed<-data.frame(zed,col)
flux$zeo<-flux$ZEMPL2010_RESID
flux$zed<-flux$ZEMPL2010_TRAV
flux<-merge(flux,dzeo,by="zeo")
flux<-merge(flux,dzed,by="zed")
## Construction de la matrice des poids
lien<-matrix(0,nrow=297,ncol=297)
for (i in 1:297)
{   for (j in 1:297)
                {ze<-flux$IPONDI[flux$lig==i & flux$col==j]
                                if(length(ze)>0)
                                lien[i,j]<-ze
                }
}
mig.w<-mat2listw(lien,style="W")
```

Modèles linéaires spatiaux : estimations complémentaires

```
### Modèle SAC
```

```
ze.sac<-sacsarlm(modele, data=donnees_ze, matrice)
summary(ze.sac)

### Modèle SLX
ze.slx<-lmSLX(modele, data=donnees_ze, matrice)
summary(ze.slx)

### Modèle SDEM
ze.sdem<-errorsarlm(modele, data=donnees_ze, matrice, etype="emixed")
summary(ze.sdem)

### Modèle Manski
ze.manski<-sacsarlm(modele, data=donnees_ze, matrice, type="sacmixed")
summary(ze.manski)
```

Annexe 2 : Modèle SDM, pour différentes matrices de voisinage

	(1) SDM Contiguïté	(2) SDM 2 Voisins	(3) SDM 5 Voisins	(4) SDM 10 Voisins	(5) SDM Distance	(6) SDM Endogène
Taux d'activité	-0.486***	-0.485***	-0.513***	-0.494***	-0.472***	-0.485***
	(0.042)	(0.042)	(0.041)	(0.041)	(0.042)	(0.042)
% Actifs Peu Diplômés	0.180***	0.215***	0.186***	0.179***	0.182***	0.184***
	(0.027)	(0.028)	(0.028)	(0.027)	(0.027)	(0.028)
% Jeunes Actifs 15-30 ans	0.196***	0.232***	0.219***	0.211***	0.209***	0.181***
	(0.047)	(0.046)	(0.047)	(0.048)	(0.046)	(0.047)
% Emploi Industriel	-0.016	-0.024**	-0.020*	-0.022*	-0.015	-0.026**
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
% Emploi Public	-0.037**	-0.038**	-0.044***	-0.048***	-0.042**	-0.050*
	(0.017)	(0.017)	(0.017)	(0.017)	(0.016)	(0.016)
$\hat{\rho}$	0.601***	0.448***	0.606***	0.647***	0.629***	0.609***
	(0.057)	(0.050)	(0.057)	(0.068)	(0.064)	(0.051)
$\hat{\theta}$, Taux d'activité	0.153**	0.094	0.209***	0.207**	0.157*	0.149**
	(0.075)	(0.057)	(0.072)	(0.087)	(0.083)	(0.075)
$\hat{\theta}$, % Actifs Peu Diplômés	-0.114***	-0.133***	-0.126***	-0.134***	-0.135***	-0.145***
	(0.040)	(0.034)	(0.041)	(0.047)	(0.045)	(0.040)
$\hat{\theta}$, % Jeunes Actifs 15-30 ans	-0.124*	-0.153***	-0.167***	-0.131*	-0.140*	-0.090
	(0.069)	(0.053)	(0.065)	(0.078)	(0.072)	(0.068)
$\hat{\theta}$, % Emploi Industriel	-0.046**	-0.029**	-0.030	-0.035	-0.044**	-0.031*
	(0.021)	(0.015)	(0.019)	(0.022)	(0.020)	(0.018)
$\hat{\theta}$, % Emploi Public	-0.029	- 0.031	-0.020	- 0.005	-0.024	0.015
	(0.033)	(0.022)	(0.031)	(0.043)	(0.037)	(0.031)
Constante	28.582***	33.848***	26.710***	24.504***	27.456***	27.662***
	(6.184)	(4.814)	(5.844)	(7.372)	(6.766)	(6.312)
Observations	297	297	297	297	297	297
AIC	968	985	970	971	960	987
Test Facteur Commun	0.002	0.001	0.040	0.035	0.004	0.000
Test LM residual auto.	0.054	0.263	0.071	0.715	0.572	0.135

TABLE 6.7 – Modèle SDM, pour différentes matrices de voisinage
Note : Le modèle SDM est estimé avec 6 matrices de voisinage différentes. Les écarts-types sont indiqués entre parenthèses. Significativité : * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Références - Chapitre 6

ABREU, Maria, Henri DE GROOT et Raymond FLORAX (2004). « Space and growth : a survey of empirical evidence and methods ».

ALDSTADT, Jared et Arthur GETIS (2006). « Using AMOEBA to create a spatial weights matrix and identify spatial clusters ». *Geographical Analysis* 38.4, p. 327–343.

ANSELIN, Luc (2001). « Spatial econometrics ». *A companion to theoretical econometrics* 310330.

— (2002a). « Under the hood : Issues in the specification and interpretation of spatial regression models ». *Agricultural economics* 27.3, p. 247–267.

ANSELIN, Luc et Daniel A GRIFFITH (1988). « Do spatial effects really matter in regression analysis ? » *Papers in Regional Science* 65.1, p. 11–34.

ANSELIN, Luc et al. (1996). « Simple diagnostic tests for spatial dependence ». *Regional science and urban economics* 26.1, p. 77–104.

ARBIA, Giuseppe (2014). *A primer for spatial econometrics : with applications in R*. Springer.

BARRIOS, Thomas et al. (2012). « Clustering, spatial correlations, and randomization inference ». *Journal of the American Statistical Association* 107.498, p. 578–591.

BECK, Nathaniel, Kristian Skrede GLEDITSCH et Kyle BEARDSLEY (2006). « Space is more than geography : Using spatial econometrics in the study of political economy ». *International studies quarterly* 50.1, p. 27–44.

BHATTACHARJEE, Arnab et Chris JENSEN-BUTLER (2013). « Estimation of the spatial weights matrix under structural constraints ». *Regional Science and Urban Economics* 43.4, p. 617–634.

BLANC, Michel et François HILD (2008). « Analyse des marchés locaux du travail : du chômage à l'emploi ». fre. *Economie et Statistique* 415.1, p. 45–60. ISSN : 0336-1454. DOI : 10.3406/estat.2008.7019. URL : https://www.persee.fr/doc/estat_0336-1454_2008_num_415_1_7019.

BRUNSDON, C. et al. (2002). « Geographically weighted summary statistics : a framework for localised exploratory data analysis ». *Computers, Environment and Urban Systems*.

BRUNSDON, Chris et Lex COMBER (2015). *An Introduction to R for Spatial Analysis Et Mapping*. Sage London.

BRUNSDON, Chris, A Stewart FOTHERINGHAM et Martin E CHARLTON (1996). « Geographically weighted regression : a method for exploring spatial nonstationarity ». *Geographical analysis* 28.4, p. 281–298.

CORRADO, Luisa et Bernard FINGLETON (2012). « Where is the economics in spatial econometrics ? » *Journal of Regional Science* 52.2, p. 210–239.

DUBIN, Robin A (1998). « Spatial autocorrelation : a primer ». *Journal of housing economics* 7.4, p. 304–327.

ELHORST, J Paul (2010). « Applied spatial econometrics : raising the bar ». *Spatial Economic Analysis* 5.1, p. 9–28.

FAFCHAMPS, Marcel (2015). « Causal Effects in Social Networks ». *Revue économique* 66.4, p. 657–686.

FINGLETON, Bernard et Julie LE GALLO (2008). « Estimating spatial models with endogenous variables, a spatial lag and spatially dependent disturbances : finite sample properties ». *Papers in Regional Science* 87.3, p. 319–339.

— (2012). « Endogeneité et autocorrélation spatiale : quelle utilité pour le modèle de Durbin ? » *Revue d'Économie Régionale & Urbaine* 1, p. 3–17.

FLACHAIRE, Emmanuel (2005). « Bootstrapping heteroskedastic regression models : wild bootstrap vs. pairs bootstrap ». *Computational Statistics & Data Analysis* 49.2, p. 361–376.

FLOCH, J.M. (2012). « Détection des disparités socio-économiques - L'apport de la statistique spatiale ». *Documents de Travail Insee*.

FLORAX, Raymond JGM, Hendrik FOLMER et Sergio J REY (2003). « Specification searches in spatial econometrics : the relevance of Hendry's methodology ». *Regional Science and Urban Economics* 33.5, p. 557–579.

GIBBONS, Stephen et Henry G OVERMAN (2012). « Mostly pointless spatial econometrics ? » *Journal of Regional Science* 52.2, p. 172–191.

GIVORD, Pauline et al. (2016). « Quels outils pour mesurer la ségrégation dans le système éducatif ? Une application à la composition sociale des collèges français ». *Education et formation*.

GRISLAIN-LETRÉMY, Céline et Arthur KATOSSKY (2013). « Les risques industriels et le prix des logements ». *Economie et statistique* 460.1, p. 79–106.

HARRIS, Richard, John MOFFAT et Victoria KRAVTSOVA (2011). « In search of 'W' ». *Spatial Economic Analysis* 6.3, p. 249–270.

KELEJIAN, Harry H et Gianfranco PIRAS (2014). « Estimation of spatial models with endogenous weighting matrices, and an application to a demand model for cigarettes ». *Regional Science and Urban Economics* 46, p. 140–149.

KELEJIAN, Harry H et Ingmar R PRUCHA (2007). « HAC estimation in a spatial framework ». *Journal of Econometrics* 140.1, p. 131–154.

— (2010a). « Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances ». *Journal of Econometrics* 157.1, p. 53–67.

KELEJIAN, H.H. et I.R. PRUSHA (2010b). « Spatial models with spatially lagged dependent variables and incomplete data ». *Journal of geographical systems*.

LE GALLO, Julie (2002). « Econométrie spatiale : l'autocorrélation spatiale dans les modèles de régression linéaire ». *Economie & prévision* 4, p. 139–157.

— (2004). « Hétérogénéité spatiale ». *Économie & prévision* 1, p. 151–172.

LEE, Lung-Fei (2004). « Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models ». *Econometrica* 72.6, p. 1899–1925.

LESAGE, James (2014). « What regional scientists need to know about spatial econometrics ».

LESAGE, James et Robert K PACE (2009). *Introduction to spatial econometrics*. Chapman et Hall/CRC.

LEUNG, Yee, Chang-Lin MEI et Wen-Xiu ZHANG (2000). « Statistical tests for spatial nonstationarity based on the geographically weighted regression model ». *Environment and Planning A* 32.1, p. 9–32.

LOONIS, Vincent (2012). « Non-réponse à l'Enquête Emploi et modèles probit spatiaux ».

LOTTMANN, Franziska (2013). « Spatial dependence in German labor markets ».

MANSKI, Charles F (1993). « Identification of Endogenous Social Effects : The Reflection Problem ». *Review of Economic Studies* 60.3, p. 531–542.

OSLAND, Liv (2010). « An application of spatial econometrics in relation to hedonic house price modeling ». *Journal of Real Estate Research* 32.3, p. 289–320.

PIRAS, Gianfranco et al. (2010). « sphet : Spatial models with heteroskedastic innovations in R ». *Journal of Statistical Software* 35.1, p. 1–21.

SLADE, Margaret E (2005). « The role of economic space in decision making ». *Annales d'Economie et de Statistique*, p. 1–20.

THOMAS-AGNAN, Christine, Thibault LAURENT et Michel GOULARD (2014). « About predictions in spatial autoregressive models ».

WAELBROECK, Patrick (2005). « The Role of Economic Space in Decision Making : Comment ». *Annales d'Economie et de Statistique*, p. 29–31.

WANG, Wei et Lung-Fei LEE (2013b). « Estimation of spatial autoregressive models with randomly missing data in the dependent variable ». *The Econometrics Journal* 16.1, p. 73–102.

WHEELER, D et A PÁEZ (2009). *Geographically weighted regression. 1er MM, Getis A (eds) Handbook of applied spatial analysis*.

7. Économétrie spatiale sur données de panel

BOUAYAD AGHA SALIMA
GAINS (TEPP) et Crest
Le Mans Université
LE GALLO JULIE
CESAER, AgroSup Dijon, INRA,
Université de Bourgogne Franche-Comté, F-21000 Dijon
VÉDRINE LIONEL
CESAER, AgroSup Dijon, INRA,
Université de Bourgogne Franche-Comté, F-21000 Dijon

Résumé

Ce chapitre propose une présentation synthétique des méthodes d'économétrie spatiale appliquées aux données de panel. Nous insistons principalement sur les spécifications et les méthodes implémentées dans le package *splm* disponible sous R. Nous illustrons notre présentation par une analyse de la deuxième "loi" de Verdoorn avant de présenter des extensions récentes des modèles spatiaux sur données de panel.

R La lecture préalable du chapitre 6 : "économétrie spatiale : modèles courants" est recommandée.

Introduction

Les données de panel concernent des observations liées à un ensemble d'individus (firmes, ménages, collectivités locales) observés à plusieurs dates (HSIAO 2014). Relativement aux données en coupe transversale, on considère que le fait de pouvoir disposer d'informations dans les dimensions individuelles et temporelles présente trois avantages principaux. Le gain d'information lié à l'exploitation de la double dimension des données permet de contrôler la présence d'hétérogénéité inobservable. La taille des échantillons généralement plus élevée permet d'améliorer la précision des estimations. Enfin, les données de panel permettent de modéliser des relations dynamiques.

Après une première génération de modèles spatiaux spécifiés pour données en coupe transversale (ELHORST 2014b), de nombreuses applications en économétrie spatiale reposent aujourd'hui sur des données de panel. En effet, si les spécifications a-spatiales sur données de panel permettent effectivement de contrôler une certaine forme d'hétérogénéité inobservée, la dépendance des coupes est prise en compte sans toujours être identifiée ou modélisée et ces modèles ne captent pas le cas particulier des effets de dépendance spatiale. De manière similaire aux modèles en coupe, l'introduction d'effets spatiaux dans les modèles en données de panel permet ainsi de mieux prendre en compte l'interdépendance entre les individus.

Dans ce chapitre, nous présentons les principales spécifications des panels spatiaux, en partant des spécifications standard en données de panel (section 7.1). La section 7.2 est consacrée à la présentation des méthodes d'estimation, et la section 7.3 décrit les principaux tests de spécifications spécifiques aux panels spatiaux. Nous proposons une application empirique en testant la deuxième loi de Verdoorn dans le cadre d'un panel de régions européennes (NUTS3) entre 1991 et 2008 (section 7.4). La section 7.5 présente quelques extensions récentes des panels spatiaux.

7.1 Spécifications

Cette section présente les principales spécifications utilisées pour les modèles statiques sur données de panel avec prise en compte des interactions spatiales. Nous ne considérons que le cas des panels cylindrés : les individus sont observés à toutes les périodes. Les travaux portant sur les méthodes d'estimation sur panels spatiaux non cylindrés sont encore peu développés. Les modèles dynamiques seront brièvement évoqués dans la section 7.5.1. Après un bref rappel de ce qui caractérise les spécifications standard sur données de panel (sans dépendance spatiale) et de ce qui distingue les effets spécifiques fixes des effets aléatoires, nous présentons les différentes façons de prendre en compte l'autocorrélation spatiale dans le contexte de ces modèles.

7.1.1 Modèle standard : modéliser les effets spécifiques individuels

Relativement aux données en coupe transversale, les données de panel, *i.e.* plusieurs observations pour les mêmes individus, permettent de tenir compte de l'influence de certaines caractéristiques non observées invariantes dans le temps de ces individus.
Pour un échantillon comportant des informations sur un ensemble d'individus indicés par $i = 1, ..., N$ que l'on suppose observables pendant toute la période d'étude $t = 1, ..., T$ (*i.e.* il n'y a ni attrition, ni observations manquantes), le modèle standard (*a-spatial*) s'écrit :

$$y_{it} = x_{it}\beta + z_i\alpha + \varepsilon_{it} \tag{7.1}$$

Les k variables explicatives du modèle sont regroupées dans k vecteurs x_{it} de dimension $(1,k)$ (qui n'inclut pas de vecteur unitaire) et sont supposées exogènes. Le vecteur β de dimension $(k,1)$

désigne le vecteur des paramètres inconnus à estimer. L'hétérogénéité, ou effet spécifique individuel, est captée par le terme $z_i\alpha$. Le vecteur z_i comprend un terme constant et un ensemble de variables spécifiques aux individus, invariantes dans le temps, qui peuvent être observées (sexe, éducation, etc.) ou non (préférences, compétences, etc.). Les hypothèses formulées sur les termes d'erreur ε_{it} dépendent du type de modèle considéré. En effet, selon la nature des variables prises en compte dans le vecteur z_i, on peut considérer trois classes de modèle : le modèle sur données empilées, le modèle à effets fixes et le modèle à effets aléatoires.

Le premier type de modèle, sur données empilées, correspond au cas pour lequel z_i ne comprend qu'une constante :

$$y_{it} = x_{it}\beta + \alpha + \varepsilon_{it} \tag{7.2}$$

où $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$. L'hétérogénéité individuelle n'est pas modélisée ; la spécification conduit à un simple empilement des données en coupes transversales. Dans ce cas un estimateur convergent et efficace de β et de α est obtenu par la méthode des Moindres Carrés Ordinaires (MCO).

Dans le second modèle, dit à effets fixes, l'hétérogénéité individuelle est modélisée par la prise en compte d'effets spécifiques individuels constants dans le temps. Ce modèle s'écrit :

$$y_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it} \tag{7.3}$$

où l'effet fixe α_i est un paramètre (moyenne conditionnelle) à estimer constant dans le temps et $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$. Dans ce modèle, les différences de comportement inobservables sont ainsi captées par ces paramètres estimables. Ce modèle est alors particulièrement adapté dès lors que l'échantillon est exhaustif au regard de la population qu'il concerne et que le modélisateur souhaite restreindre les résultats obtenus à l'échantillon qui a permis de les obtenir. Les effets individuels α_i peuvent être corrélés avec les variables explicatives x_{it} et l'estimateur *within* (*i.e.* l'estimateur des MCO obtenu à partir d'un modèle où les variables explicatives et expliquée sont centrées sur leur moyenne individuelle respective, voir équation 7.20) reste convergent.

Dans le troisième modèle, à effets aléatoires, l'hétérogénéité individuelle est modélisée par la prise en compte d'effets spécifiques individuels aléatoires (constants au cours du temps). On fait l'hypothèse que cette hétérogénéité individuelle inobservable n'est pas corrélée avec x_{it} :

$$\begin{aligned} y_{it} &= x_{it}\beta + \alpha + u_{it} \\ u_{it} &= \alpha_i + \varepsilon_{it} \end{aligned} \tag{7.4}$$

où $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$.
Contrairement au modèle à effets fixes, les effets individuels ne sont plus des paramètres à estimer, mais les réalisations d'une variable aléatoire. Ce modèle est donc adapté si les spécificités individuelles sont reliées à des causes aléatoires. Il est également préférable au modèle à effets fixes lorsque les individus présents dans l'échantillon sont tirés d'une population plus large et que l'objectif de l'étude empirique est de généraliser les résultats obtenus à la population. Ce modèle présente l'avantage de fournir des estimations plus précises que celles obtenues à partir du modèle à effets fixes. Il s'estime usuellement à l'aide de la méthode des Moindres Carrés Généralisés (MCG).

Dans la suite de ce chapitre, nous adoptons une présentation générale de la spécification de la nature des effets individuels en distinguant les effets individuels fixes des effets aléatoires. Nous présentons également les tests usuels de spécification permettant de choisir la bonne méthode d'estimation et donc la spécification la plus adaptée pour modéliser l'hétérogénéité. Cependant, si ces modèles permettent de prendre en compte l'hétérogénéité individuelle, ils ont en commun

avec le modèle standard en coupe transversale de reposer sur l'hypothèse que les individus sont indépendants les uns des autres. Si les données portent sur des individus pour lesquels on dispose d'informations géolocalisées et que l'on suppose l'existence d'interactions spatiales, cette hypothèse n'est plus acceptable. Il convient donc d'étendre les spécifications présentées précédemment en prenant en compte l'autocorrélation spatiale.

7.1.2 Les effets spatiaux dans les modèles en données de panel

Comme pour les modèles en coupe transversale, la prise en compte de l'autocorrélation spatiale peut se faire de plusieurs manières : par des variables spatiales décalées, endogènes ou exogènes, ou par une autocorrélation spatiale des erreurs.

Les effets spatiaux dans les modèles sur données empilées

Dans un premier temps, nous reprenons le modèle empilé en incorporant ces trois termes spatiaux potentiels :

$$\begin{aligned} y_{it} &= \rho \sum_{i \neq j} w_{ij} y_{jt} + x_{it}\beta + \sum_{i \neq j} w_{ij} x_{jt}\theta + \alpha + u_{it} \\ u_{it} &= \lambda \sum_{i \neq j} w_{ij} u_{jt} + \varepsilon_{it} \end{aligned} \tag{7.5}$$

w_{ij} est un élément d'une matrice de pondération spatiale W_N de dimension (N,N) dans laquelle sont définies les relations de voisinage entre les individus de l'échantillon. Par convention, les éléments diagonaux w_{ii} sont tous fixés à zéro. La matrice de poids est généralement normalisée en ligne. La plupart des travaux académiques considèrent une matrice de pondération spatiale fixe dans le temps. La variable $\sum_{i \neq j} w_{ij} y_{jt}$ désigne la variable endogène spatialement décalée ; elle est égale à la valeur moyenne de la variable dépendante prise par les voisins (au sens de la matrice de poids) de l'observation i. Le paramètre ρ capte l'effet d'interaction endogène. L'interaction spatiale est également prise en compte par la spécification d'un processus autorégressif spatial dans les erreurs $\sum_{i \neq j} u_{jt}$ selon lequel les chocs inobservables affectant l'individu i interagissent avec les chocs affectant son voisinage. Le paramètre λ capte un effet corrélé des inobservables. Enfin, un effet contextuel (ou d'interaction exogène) est capté par le vecteur θ de dimension $(k,1)$. Comme précédemment, on suppose que $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$.

En empilant les données pour chaque période t, le modèle précédent s'écrit de la façon suivante :

$$\begin{aligned} y_t &= \rho W_N y_t + x_t\beta + W_N x_t\theta + \alpha + u_t \\ u_t &= \lambda W_N u_t + \varepsilon_t \end{aligned} \tag{7.6}$$

où y_t est le vecteur de dimension $(N,1)$ des observations de la variable expliquée pour la période t, x_t est la matrice (N,k) des observations sur les variables explicatives pour la période t. Enfin, en empilant les données pour tous les individus, le modèle s'écrit sous forme matricielle de la façon suivante :

$$\begin{aligned} y &= \rho (I_T \otimes W_N) y + x\beta + (I_T \otimes W_N) x\theta + \alpha + u \\ u &= \lambda (I_T \otimes W_N) u + \varepsilon \end{aligned} \tag{7.7}$$

où $\otimes$ désigne le produit Kronecker et $(I_T \otimes W_N)$ est une matrice de dimension (NT,NT) de la forme suivante :

$$\begin{pmatrix} W_N & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & W_N \end{pmatrix}$$

Comme vu dans le chapitre précédent : "économétrie spatiale : modèles courants", les paramètres de ce modèle ne sont généralement pas identifiables (MANSKI 1993). Il convient de faire des choix sur la nature des termes spatiaux à privilégier dans le modèle. Ces choix peuvent s'appuyer sur une modélisation théorique et/ou reposer sur une stratégie de spécification allant du spécifique au général à partir des résultats des tests du multiplicateur de Lagrange utilisés pour les modèles en coupe transversale.

Cependant, l'intérêt du modèle sur données empilées reste limité, puisque celui-ci ne permet pas de considérer la présence d'hétérogénéité individuelle alors que les individus sont susceptibles de différer du fait de caractéristiques inobservables ou difficilement mesurables. Selon la manière dont est modélisée l'hétérogénéité inobservable (fixe par opposition à aléatoire) l'omission de ces caractéristiques peut compromettre la convergence des estimateurs pour les paramètres β, θ et α. En conséquence, les modèles à effets spécifiques, fixes ou aléatoires, sont à privilégier. Nous présentons, dans ce cadre, les spécifications faisant intervenir simultanément un ou deux des termes spatiaux présentés plus haut, pour lesquels nous disposons d'estimateurs documentés dans la littérature.

Les effets spatiaux dans les modèles à effets fixes

Plusieurs spécifications spatiales peuvent être considérées pour tenir compte de l'autocorrélation spatiale dans le modèle à effets fixes. La première spécification est le modèle autorégressif spatial (SAR) , qui s'écrit :

$$y_{it} = \rho \sum_{i \neq j} w_{ij} y_{jt} + x_{it}\beta + \alpha_i + u_{it} \tag{7.8}$$

où $u_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. L'interaction spatiale est ici modélisée à travers l'introduction de la variable dépendante spatialement décalée ($\sum_{i \neq j} w_{ij} y_{jt}$). Comme dans les modèles en coupe transversale, l'introduction de cette variable implique des effets de débordement globaux : en moyenne, la valeur de y au temps t pour une observation i n'est pas seulement expliquée par les valeurs des variables explicatives pour cette observation, mais aussi par celles associées à toutes les observations (voisines de i ou non). C'est l'effet de multiplicateur spatial. Un effet global de diffusion spatiale est également à l'œuvre : un choc aléatoire dans une observation i au temps t affecte non seulement la valeur de y de cette observation à la même période mais a également un effet sur les valeurs de y des autres observations.

Le deuxième modèle est connu sous le nom de modèle à erreur spatiale (SEM) :

$$\begin{aligned} y_{it} &= x_{it}\beta + \alpha_i + u_{it} \\ u_{it} &= \lambda \sum_{i \neq j} w_{ij} u_{jt} + \varepsilon_{it} \end{aligned} \tag{7.9}$$

où $u_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. L'interaction spatiale est captée à travers une spécification autorégressive spatiale du terme d'erreur ($\lambda \sum_{i \neq j} w_{ij} u_{jt}$). Seul l'effet de diffusion spatiale est présent dans un modèle SEM, il reste cependant global.

Un troisième modèle, préconisé par LESAGE et al. 2009, est le modèle spatial de Durbin (SDM) qui contient une variable dépendante spatialement décalée ($\sum_{i \neq j} w_{ij} y_{jt}$) et des variables explicatives spatialement décalées ($\sum_{i \neq j} w_{ij} x_{jt}$) :

$$y_{it} = \rho \sum_{i \neq j} w_{ij} y_{jt} + x_{it}\beta + \sum_{i \neq j} w_{ij} x_{jt}\theta + \alpha_i + u_{it} \tag{7.10}$$

où $u_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$.

Une alternative à ce modèle est le modèle de Durbin spatial dans les erreurs (SDEM) , qui est composé d'un terme d'erreur spatialement autocorrélé ($\sum_{i\neq j} w_{ij}u_{jt}$) et des variables explicatives spatialement décalées ($\sum_{i\neq j} w_{ij}x_{jt}$) :

$$\begin{aligned} y_{it} &= x_{it}\beta + \sum_{i\neq j} w_{ij}x_{jt}\theta + \alpha_i + u_{it} \\ u_{it} &= \lambda \sum_{i\neq j} w_{ij}u_{jt} + \varepsilon_{it} \end{aligned} \tag{7.11}$$

où $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$. À travers l'autocorrélation spatiale des erreurs, il existe bien un effet de diffusion globale mais il n'y a pas d'effet multiplicateur. En effet, introduire des variables spatiales explicatives décalées induit des effets de débordements locaux et non globaux (voir chapitre 6 : "économétrie spatiale : modèle courants").

Enfin, certains auteurs utilisent une modélisation faisant intervenir simultanément un processus autorégressif spatial de la variable dépendante et du terme d'erreur (SARAR), les pondérations spatiales (w_{ij} et m_{ij}) étant distinctes pour chacun des processus (LEE et al. 2010b ; ERTUR et al. 2015) :

$$\begin{aligned} y_{it} &= \rho \sum_{i\neq j} w_{ij}y_{jt} + x_{it}\beta + \alpha_i + u_{it} \\ u_{it} &= \lambda \sum_{i\neq j} m_{ij}u_{jt} + \varepsilon_{it} \end{aligned} \tag{7.12}$$

où $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$.

Les effets spatiaux dans les modèles à effets aléatoires

Dans les modèles à effets aléatoires, les effets individuels non observés sont supposés non corrélés avec les autres variables explicatives du modèle et peuvent donc être traités comme des composants du terme d'erreur. Dans ce contexte, le modèle SAR s'écrit de manière similaire à ce qui a été proposé dans le cadre du modèle à effets fixes, à l'exception du terme d'effet individuel :

$$\begin{aligned} y_{it} &= \rho \sum_{i\neq j} w_{ij}y_{jt} + x_{it}\beta + \alpha + u_{it} \\ u_{it} &= \alpha_i + \varepsilon_{it} \end{aligned} \tag{7.13}$$

où $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$.

L'effet aléatoire étant une partie du terme d'erreur, deux spécifications SEM sont proposées dans la littérature. Dans la première (SEM-RE), l'effet de diffusion spatiale n'est considéré que pour le terme d'erreur idiosyncratique [1] et non pour l'effet individuel aléatoire (BALTAGI et al. 2003). On peut donc écrire :

$$\begin{aligned} y_{it} &= x_{it}\beta + u_{it} \\ u_{it} &= \alpha_i + \lambda \sum_{i\neq j} w_{ij}u_{jt} + v_{it} \end{aligned} \tag{7.14}$$

où $v_{it} \overset{i.i.d.}{\sim} \mathcal{N}iid(0,\sigma^2)$.

1. ie. le terme d'erreur individuel temporel.

Dans une seconde spécification (RE-SEM), suggérée par KAPOOR et al. 2007 (on désigne souvent cette spécification par KKP), on considère que la structure de corrélation spatiale s'applique à la fois aux effets individuels et à la composante restante du terme d'erreur :

$$\begin{aligned} y_{it} &= x_{it}\beta + \alpha + u_{it} \\ u_{it} &= \lambda \sum_{i \neq j} w_{ij} u_{jt} + v_{it} \\ v_{it} &= \alpha_i + \varepsilon_{it} \end{aligned} \tag{7.15}$$

où $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.

Ces deux spécifications impliquent des effets de reports spatiaux assez différents régis par des matrices de variances-covariances de structure différente, ce qui a des implications en matière d'estimation. D'autre part, comme le soulignent BALTAGI et al. 2013, ces deux modèles ont des implications différentes : dans le premier, seule la composante qui varie dans le temps se diffuse spatialement, tandis que dans le second cela caractérise également la composante permanente.

Enfin, on peut finalement envisager une spécification plus générale comme celle suggérée par BALTAGI et al. 2007 [2] :

$$\begin{aligned} y_{it} &= x_{it}\beta + u_{it} \\ u_{it} &= \alpha_i + \lambda \sum_{i \neq j} w_{ij} u_{jt} + v_{it} \\ \alpha_i &= \eta \sum_{i \neq j} w_{ij} \alpha_j + e_i \end{aligned} \tag{7.16}$$

où $e_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.

Le processus autorégressif spatial sur l'effet individuel s'interprète comme un effet de diffusion spatiale permanent sur la période.

7.1.3 Interprétation des coefficients en présence d'un terme autorégressif spatial

Comme dans les modèles de régression en coupe transversale, on peut, à partir des spécifications précédentes, donner l'expression des effets marginaux des variables explicatives ainsi que celles des impacts directs, indirects et totaux qui facilitent l'interprétation des coefficients des modèle estimés. En effet, à la différence des modèles a-spatiaux, l'effet marginal d'une variation d'une variable explicative peut être différent d'un individu à l'autre. En effet, du fait des interactions spatiales, la variation d'une variable explicative pour un individu affecte directement son résultat et indirectement les résultats de tous les autres individus. La fonction `impacts.splm`, du package *splm* de R, étend les méthodes de calcul d'impact développées pour les modèles en coupe en tenant compte de la spécificité de la dimension (NT, NT) de la matrice de pondération spatiale qui intervient dans les spécifications sur données de panel [3].

Quelle que soit la nature des données prises en compte, du fait des interactions spatiales, toute variation d'une variable explicative x_k pour un individu i entraîne une variation de la variable dépendante pour ce même individu (effet direct) mais également pour les autres (effet indirect). Pour une même variation unitaire ces effets peuvent être différents d'un individu à l'autre. Les mesures d'impacts proposées par LESAGE et al. 2009 sont donc des effets moyens dont l'expression va dépendre de la spécification spatiale retenue.

2. Ce modèle admet la spécification de KAPOOR et al. 2007 comme cas particulier pour $\eta = \lambda$ et BALTAGI et al. 2003 pour $\eta = 0$.

3. Le lecteur peut se référer à PIRAS 2014 pour plus de détails dans le calcul des effets directs, indirects et totaux sous R.

Dans le modèle de régression en coupe, si l'on part de la forme réduite du modèle autorégressif spatial (SAR), les mesures d'impacts de la variable explicative k se déduisent de l'équation suivante :

$$S_k(W_N) = (I_N - \lambda W_N)^{-1} I_N \beta_k. \tag{7.17}$$

Par analogie, dans un panel spatial statique, pour calculer les effets directs et indirects il suffit de remplacer W_N, invariante dans le temps, par la matrice diagonale par blocs $W_N = I_N \otimes W_N$. C'est cette matrice qui figure sur la diagonale de W_N dans l'équation précédente (PIRAS 2014), soit :

$$S_k(I_N \otimes W_N) = (I_{NT} - \lambda (I_N \otimes W))^{-1} I_{NT} \beta_k. \tag{7.18}$$

Plus généralement, si l'on considère un modèle spatial Durbin (SDM ; équation 7.10), la matrice des dérivées partielles de la variable dépendante, pour chacune des unités, relativement à la variable explicative k à un instant t donné s'écrit :

$$\Gamma = \left(\frac{\partial y}{\partial x_{1k}} \dots \frac{\partial y}{\partial x_{Nk}} \right)_t = (I - \rho W_N)^{-1} \begin{pmatrix} \beta_k & w_{12}\theta_k & \dots & w_{1N}\theta_k \\ w_{21}\theta_k & \beta_k & \dots & w_{2N}\theta_k \\ \vdots & \vdots & \vdots & \vdots \\ w_{N1}\theta_k & w_{N2}\theta_k & \dots & \beta_k \end{pmatrix}. \tag{7.19}$$

LESAGE et al. 2009 définissent l'effet direct comme la moyenne des éléments diagonaux de la matrice figurant dans le terme de droite de l'équation 7.19 et l'effet indirect comme la moyenne de la somme des éléments en lignes (ou en colonnes) en dehors de ceux situés sur la diagonale principale.

Dans le cas du modèle SEM, la matrice du terme de droite de l'équation 7.19 est une matrice diagonale dont les éléments sont égaux à β_k. De ce fait, l'effet direct d'une variation de la variable explicative k est égal à β_k et l'effet indirect est nul comme dans les modèles a-spatiaux et dans les modèles spatiaux en coupe.

Dans le cas du modèle SAR, bien que les éléments en dehors de la diagonale principale de la seconde matrice du terme de droite de l'équation 7.19 soient nuls, du fait de la dimension de W, le calcul des effets directs et indirects exige de mettre en œuvre des calculs matriciels et de calculer la trace de la matrice Γ qui fait intervenir des puissances de W. D'autre part, les statistiques permettant de tester la significativité de ces mesures d'impact sont obtenues par simulation de Monte Carlo (pour plus de détail voir PIRAS 2014).

7.2 Méthodes d'estimations

Deux grandes catégories de méthodes d'estimation des modèles spatiaux sur données de panel sont principalement utilisées : les méthodes fondées sur le principe du maximum de vraisemblance et les méthodes fondées sur la méthode des moments généralisée (incluant les variables instrumentales). Comme précédemment, nous nous restreignons ici au cas standard d'un panel cylindré et d'une matrice de pondération spatiale fixe dans le temps. Généralement, les estimateurs par le maximum de vraisemblance (MV) sont plus efficaces, mais reposent sur des conditions plus fortes sur la distribution du terme d'erreur. La méthode des moments généralisée (MMG) est souvent privilégiée car moins coûteuse en temps de calcul et plus facile à mettre en œuvre. D'autre part, dans la majorité des cas, comme ces estimateurs ne reposent pas sur l'hypothèse de normalité, les estimateurs que cette méthode permet d'obtenir sont plus robustes à l'hétéroscédasticité. Enfin, la flexibilité que permet la définition des conditions sur les moments permet également d'estimer les modèles spatiaux en présence d'une variable explicatives endogène. Ces deux méthodes sont implémentables sous R. Cette section présente les estimateurs des modèles à effets fixes (section 7.3.1), puis des modèles à effets aléatoires (section 7.3.2).

7.2.1 Modèle à effets fixes

Encadré 7.2.1 — Estimation d'un modèle à effets fixes par maximum de vraisemblance. Lorsque l'effet individuel spécifique est considéré comme fixe, la procédure la plus souvent employée (approche directe) consiste à transformer les variables du modèle de telle sorte à éliminer l'effet fixe, puis à estimer directement le modèle sur ces variables transformées. La transformation la plus courante est la déviation intra-individuelle (*within*). Elle consiste à différencier chaque variable par rapport à sa moyenne intra-individuelle :

$$y_{it}^* = y_{it} - \frac{1}{T}\sum_{t=1}^{T} y_{it} \quad et \quad x_{it}^* = x_{it} - \frac{1}{T}\sum_{t=1}^{T} x_{it} \tag{7.20}$$

Dans un deuxième temps, l'estimation se fait à partir des variables transformées. Dans un modèle sans autocorrélation spatiale, la fonction de vraisemblance s'écrit :

$$LogL = -\frac{NT}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}^* - x_{it}^*\beta)^2 \tag{7.21}$$

Si le modèle intègre une variable endogène décalée ($\sum_{i\neq j} w_{ij}y_{jt}$), alors la fonction de vraisemblance doit être dérivée en prenant en compte l'endogénéité de $\sum_{i\neq j} w_{ij}y_{jt}$ *via* un terme jacobien (ANSELIN et al. 2006) :

$$LogL = -\frac{NT}{2}log(2\pi\sigma^2) + Tlog|I_n - \rho W| - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}^* - \rho\sum_{j\neq i} w_{ij}y_{jt}^* - x_{it}^*\beta)^2 \tag{7.22}$$

Cette fonction est très proche de celle dérivée pour le modèle SAR en coupe transversale. Son estimation suit donc une procédure semblable. Les estimateurs de β et σ^2 étant fonction de ρ, ELHORST 2003 propose d'utiliser une fonction de log vraisemblance concentrée que l'on peut maximiser à partir des résidus (u_0^* et u_1^*) de deux régressions respectivement de y_{it}^* et de $\sum_{i\neq j} w_{ij}y_{jt}^*$ sur x_{it}^* :

$$LogL_C = C + Tlog|I_n - \rho W| - \frac{NT}{2}((u_0^* - \rho u_1^*)'(u_0^* - \rho u_1^*)) \tag{7.23}$$

Il faut utiliser une procédure par itération nécessitant de fixer initialement ρ pour calculer $\hat{\beta}$ et $\hat{\sigma}^2$. Dans un deuxième temps, il faut estimer $\hat{\rho}$ de telle sorte à maximiser la fonction de log vraisemblance concentrée et recommencer à calculer $\hat{\beta}$ et $\hat{\sigma}^2$ en fixant $\hat{\rho}$ jusqu'à obtenir des résultats qui convergent numériquement.

La modélisation de l'autocorrélation spatiale à travers un terme d'erreur spatialement autocorrélé modifie seulement l'estimation de σ^2 (l'estimation de β n'est pas affectée). La méthode des moindres carrés généralisée pourrait permettre d'obtenir un estimateur de σ^2 si λ était connu. En toute généralité, ce n'est pas le cas et il est nécessaire encore une fois d'estimer de manière itérative β, λ puis σ^2. La fonction de vraisemblance concentrée peut être maximisée à l'aide des résidus (ε_{it}^*) de la régression de y_{it}^* sur x_{it}^* :

$$LogL_C = Tlog(I_N - \lambda W) - \frac{NT}{2}log(\varepsilon_{it}^*(I_N - \lambda W)'\varepsilon_{it}^*(I_N - \lambda W)) \tag{7.24}$$

LEE et al. 2010b ont remis en cause cette approche en montrant qu'elle ne permettait pas nécessairement d'obtenir des estimateurs convergents des coefficients et des écart-types.

L'ampleur des biais et les paramètres affectés diffèrent en fonction des cas. Par exemple, lorsque le modèle contient un effet fixe individuel, σ^2 est biaisé pour N grand et T fixe. Si le modèle intègre à la fois des effets temporels et individuels, les β et σ^2 seront biaisés pour N et T grand. À partir de ces résultats, LEE et al. 2010b suggèrent des corrections, spécifiques à chaque cas, permettant d'obtenir des estimateurs convergents à partir de l'approche directe. Ces corrections sont disponibles dans les principaux logiciels d'économétrie. Nous renvoyons à LEE et al. 2010b et ELHORST 2014b pour plus de précisions sur cette approche.

Encadré 7.2.2 — Estimation d'un modèle à effets fixes par la méthode des moments généralisée. Une stratégie d'estimation alternative repose sur la méthode des moments généralisée. Dans le cadre des modèles spatiaux, la stratégie proposée par KELEJIAN et al. 1999 pour les données en coupe est étendue aux données de panel par KAPOOR et al. 2007 et MUTL et al. 2011.

Pour un modèle SAR, la stratégie d'estimation mise en œuvre repose sur la méthode des variables instrumentales proposée par KELEJIAN et al. 1998 sur le modèle en déviation intra-individuelle (*within*). Les instruments utilisés sont les variables exogènes du modèle ainsi que leur décalage spatial.

Dans le cas d'un modèle SEM, la stratégie d'estimation du paramètre d'autocorrélation spatiale sur les erreurs repose sur les trois conditions sur les moments proposées par KELEJIAN et al. 1999 pour les données en coupe, celles-ci étant étendues aux résidus du modèle en déviation intra-individuelle. Les autres paramètres du modèle peuvent alors être estimés par les moindres carrés ordinaires à partir d'un modèle auquel il a été appliqué une transformation de type Cochrane-Orcutt.

7.2.2 Modèle à effets aléatoires

Encadré 7.2.3 — Estimation d'un modèle à effets aléatoires par maximum de vraisemblance. Lorsqu'on considère un modèle à effets aléatoires, on fait l'hypothèse que les effets individuels non observés ne sont pas corrélés avec les variables explicatives du modèle. Comme dans le cas du modèle à effets fixes, on peut mettre en œuvre une méthode en deux étapes en utilisant des variables pour lesquelles la transformation dépend de ϕ tel que $\phi^2 = \sigma^2/(T\sigma_\alpha^2 + \sigma^2)$, soit :

$$y_{it}^{o} = y_{it} - (1-\phi)\frac{1}{T}\sum_{t=1}^{T} y_{it} \quad et \quad x_{it}^{o} = x_{it} - (1-\phi)\frac{1}{T}\sum_{t=1}^{T} x_{it} \tag{7.25}$$

On remarque que si $\phi = 0$, on se ramène à la transformation *within* et le modèle à effets aléatoires se ramène à un modèle à effets fixes.
Dans un modèle sans autocorrélation spatiale, la fonction de vraisemblance s'écrit :

$$LogL = -\frac{NT}{2}log(2\pi\sigma^2) + \frac{N}{2}log(\phi^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}^{o} - x_{it}^{o}\beta)^2 \tag{7.26}$$

Si le modèle intègre une variable endogène décalée, alors la fonction de vraisemblance

s'écrit :

$$LogL = -\frac{NT}{2}log(2\pi\sigma^2) + Tlog|I_n - \rho W| + \frac{N}{2}log(\phi^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}^{o} - \rho\sum_{j\neq i} w_{ij}y_{jt}^{o} - x_{it}^{o}\beta)^2 \tag{7.27}$$

Pour ϕ donné, cette fonction est très proche de celle dérivée pour le modèle SAR à effets fixes. Son estimation suit donc une procédure analogue, en utilisant une log vraisemblance concentrée que l'on peut maximiser à partir des résidus $e^{o}(\phi)$ de la régression de y_{it}^{o} sur $\sum_{i\neq j} w_{ij}y_{jt}^{o}$ et x_{it}^{o} :

$$LogL_C = -\frac{NT}{2}log\left[(e^{o}(\phi))'(e^{o}(\phi))\right] + \frac{N}{2}log(\phi^2) \tag{7.28}$$

De la même manière que précédemment, il faut fixer des valeurs initiales des paramètres inconnus puis utiliser une procédure itérative jusqu'à obtenir des résultats qui convergent numériquement.

Dans le cas d'un modèle avec erreur spatialement autocorrélée (SEM), l'écriture la plus générale de la vraisemblance est assez complexe (ELHORST 2014b) et la méthode de résolution mise en œuvre dépend de la forme de la matrice de variances-covariances des erreurs qui découle de l'hypothèse formulée sur la structure de corrélation spatiale des erreurs.

Dans le cadre de la spécification SEM-RE (seul le terme d'erreur idiosyncratique est spatialement corrélé) la vraisemblance s'écrit :

$$LogL = -\frac{NT}{2}log(2\pi\sigma^2) - \frac{1}{2}log|V| + (T-1)\sum_{i=1}^{N}log|B| - \frac{1}{2\sigma^2}e'(\bar{J}_T \otimes V^{-1})e - \frac{1}{2\sigma^2}e'(E_T \otimes (B'B))e \tag{7.29}$$

où $V = T\phi' I_N + (B'B)^{-1}$, $\quad e = y - x\beta$, $\quad B = (I_N - \lambda W)$, $\quad \phi' = \frac{\sigma^2}{\sigma_\alpha}$
avec $J_T = i_T i_T^{'}$ une matrice (T,T) de 1, $\quad \bar{J}_T = \frac{J_T}{T}$, $\quad E_T = I_T - \bar{J}_T$

Compte tenu de cette structure complexe, l'algorithme de filtrage spatial suggéré par ELHORST 2003 est particulièrement adapté à la spécification dans laquelle le terme autorégressif spatial affecte la totalité du terme d'erreur. Dans le cadre de la spécification considérée par KAPOOR et al. 2007 (KKP), la matrice de variance covariance a une forme spécifique plus simple que dans le cas précédent ce qui facilite considérablement la mise en œuvre de l'estimation par le MV en deux étapes (MILLO et al. 2012).

Cette même procédure peut être mise en œuvre pour de nombreuses autres spécifications mixant des hypothèses sur la structure d'autocorrélation spatiale. Ces méthodes d'estimation sont implémentées *via* la fonction `spreml` qui permet d'estimer par le MV plus de spécifications que la fonction `spml` (MILLO 2014).

Encadré 7.2.4 — Estimation d'un modèle à effets aléatoires par la méthode des moments généralisée. Comme dans le modèle à effets fixes, la mise en œuvre de l'estimation par la méthode des moments généralisée repose sur la stratégie proposée par KELEJIAN et al. 1999 pour les données en coupe et étendue aux données de panel par KAPOOR et al. 2007 et MUTL

et al. 2011. Par exemple, dans le modèle SEM-RE, afin d'estimer le paramètre autorégressif λ et les variances des termes d'erreurs $\sigma_1^2 = \sigma_v^2 + T\sigma_\alpha^2$ et σ_v^2, ils définissent un ensemble de 6 conditions sur les moments. MILLO et al. 2012 détaillent les différentes variantes de cet estimateur selon les conditions formulées sur les moments. Ensuite, pour les paramètres du modèle, un estimateur des moindres carrés généralisés réalisables est défini basé sur une transformation de type Cochrane-Orcutt du modèle initial.

7.3 Tests de spécification

Nous présentons dans un premier temps le test de spécification d'Hausman qui permet d'arbitrer entre un modèle où les effets individuels ne sont pas corrélés avec les variables explicatives et un modèle où une telle corrélation existe. Ce test permet de définir quelle méthode d'estimation retenir. Dans un deuxième temps, nous présentons les autres tests de spécification permettant de choisir la spécification la plus adéquate.

7.3.1 Choisir entre effet fixe et effet aléatoire

Pour que le modèle à effets aléatoires soit valide, une hypothèse cruciale est que les caractéristiques inobservables ne soient pas corrélées avec les variables explicatives observables. L'hypothèse nulle du test peut se mettre sous la forme générale $\mathbb{E}[\alpha|X] = 0$. Si cette hypothèse n'est pas rejetée, les deux estimateurs MCG et *within* seront convergents. Dans le cas contraire, l'estimateur MCG ne sera pas convergent alors que l'estimateur *within* restera convergent.

Le test de spécification d'Hausman (HAUSMAN 1978) peut s'appliquer pour tester le modèle à effets aléatoires contre le modèle à effets fixes. Dans notre cas, ce test se construit en mesurant l'écart (pondéré par une matrice de variance covariance) entre les estimations produites par les estimateurs *within* (modèle à effets fixes) et MCG (modèle à effets aléatoires) dont on sait que l'un des deux (*within*) est convergent quelle que soit l'hypothèse faite sur la corrélation entre variables et caractéristiques inobservables tandis que l'autre (MCG) n'est pas convergent dans le seul cas où cette hypothèse n'est pas vérifiée. Par conséquent, une différence significative des deux estimations implique une mauvaise spécification du modèle à effet aléatoire.

MUTL et al. 2011 ont montré que ces propriétés restent valides dans un cadre spatial si l'on remplace chaque estimateur *within* et MCG par son "analogue" spatial (prenant en compte les termes d'autocorrélation spatiale). Le test d'Hausman robuste à l'autocorrélation spatiale s'écrit :

$$S_{hausman} = NT(\hat{\beta}_{MCG} - \hat{\beta}_{within})'(\hat{\sum}_{within} - \hat{\sum}_{MCG})^{-1}(\hat{\beta}_{MCG} - \hat{\beta}_{within}) \tag{7.30}$$

où $\hat{\beta}_{MCG}$ et $\hat{\beta}_{within}$ sont les estimations des paramètres obtenus respectivement par MCG et *within*, $\hat{\sum}_{within}$ et $\hat{\sum}_{MCG}$ correspondent aux éléments des matrices de variances-covariances des deux estimations.

7.3.2 Tests de spécification des effets spatiaux

Il s'agit ici de présenter certains des tests permettant de retenir la spécification la plus adéquate de la prise en compte de la dépendance spatiale. Nous insistons sur les tests mis en œuvre dans la package R *splm*. Les tests de spécification de l'autocorrélation spatiale les plus couramment utilisés reposent sur le test du multiplicateur de Lagrange. Ils permettent de tester l'absence de chacun des termes spatiaux sans avoir à estimer le modèle non contraint. Un ensemble de tests a été développé par DEBARSY et al. 2010 dans le cadre d'un modèle à effets fixes.

On complète généralement ces deux tests par leur version robuste à la forme alternative de prise en compte de l'autocorrélation spatiale. Dans ce cas, il s'agit pour le RLMlag de tester l'absence de terme autorégressif spatial lorsque le modèle contient déjà un terme autorégressif spatial dans

les erreurs (RLMlag), ou inversement pour RLMerr de tester l'absence de terme autorégressif spatial dans les erreurs lorsque le modèle contient un terme autorégressif spatial. L'interprétation des résultats de ces tests est similaire à celle présentée dans le chapitre 6 "économétrie spatiale : modèles courants" sur les données en coupe.

BALTAGI et al. 2003 et BALTAGI et al. 2007 dérivent un ensemble de tests pour toutes les combinaisons d'effets aléatoires et d'autocorrélation spatiale dans les erreurs. Ces tests ont été complétés par BALTAGI et al. 2008 qui proposent un test joint d'absence de terme autorégressif spatial en présence d'effets individuels aléatoires. Les hypothèses de ces tests, également fondés sur le principe du multiplicateur de Lagrange, sont décrites dans la table 7.1.

Test	hypothèse nulle	hypothèse alternative
LMjoint	$\lambda = \sigma_\alpha^2 = 0$	$\lambda \neq 0$ ou $\sigma_\alpha^2 \neq 0$
SLM1	$\sigma_\alpha^2 = 0$ en posant $\lambda = 0$	$\sigma_\alpha^2 \neq 0$ en posant $\lambda = 0$
SLM2	$\lambda = 0$ en posant $\sigma_\alpha^2 = 0$	$\lambda \neq 0$ en posant $\sigma_\alpha^2 = 0$
CLMerr	$\lambda = 0$ en posant $\sigma_\alpha^2 >= 0$	$\lambda \neq 0$ en posant $\sigma_\alpha^2 >= 0$
CLMrandom	$\sigma_\alpha^2 = 0$ en posant $\lambda >= 0$	$\sigma_\alpha^2 \neq 0$ en posant $\lambda >= 0$

TABLE 7.1 – Test d'autocorrélation spatiale en présence d'effet aléatoire et/ou de corrélation sérielle

Enfin, comme dans les modèles en coupe transversale, il est possible de mettre en œuvre des tests de significativité sur les coefficients dans la mesure où certains des modèles présentés précédemment présentent un caractère emboîté. Ainsi, il est possible de retrouver le modèle SAR et le modèle SEM à partir du modèle SDM avec les contraintes testables suivantes sur les paramètres, respectivement $H_0 : \theta = 0$ (test de significativité du vecteur de paramètres θ) et $H_0 : \rho\beta - \theta = 0$ (test du facteur commun). De même à partir du modèle SDEM, on retrouve le modèle SEM si l'hypothèse $H_0 : \theta = 0$ ne peut pas être rejetée.

7.4 Application empirique

7.4.1 Le modèle

Notre application empirique porte sur la deuxième "loi" de VERDOORN 1949. Cette loi relie, de manière linéaire, les taux de croissance de la productivité du travail p à ceux de l'output q dans le secteur manufacturier pour un ensemble d'économies. La spécification de base est donnée par :

$$p_{it} = b_0 + b_1 q_{it} + \varepsilon_{it} \tag{7.31}$$

où b_0 et b_1 sont les paramètres inconnus à estimer et ε_{it} est un terme d'erreur pour lequel nous supposons dans un premier temps que $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. Le paramètre b_1 est appelé le coefficient de Verdoorn pour lequel une valeur positive traduit la présence de rendements croissants (FINGLETON et al. 1998). Cette spécification a été affinée par FINGLETON 2000, 2001 afin de caractériser l'endogénéité du progrès technique. Il suppose notamment un changement technique proportionnel à l'accumulation du capital par tête et une croissance du capital par tête égale à la croissance de la productivité et des effets de débordements géographiques, liés notamment à la diffusion des technologies et du capital humain entre unités spatiales. La spécification étendue de Verdoorn qui

découle de ces analyses est [4] :

$$p_{it} = b_0 + b_1 q_{it} + b_2 G_{it} + b_3 u_{it} + b_4 d_{it} + \varepsilon_{it} \tag{7.32}$$

où G correspond à l'écart technologique (approché par le différentiel de productivité du travail) au début de la période entre chaque unité et l'unité spatiale "leader". Dans le cadre des modèles de croissance endogène, les unités spatiales avec un retard technologique sont susceptibles de connaître une croissance de la productivité plus faible que celle des unités spatiales plus développées. u est une mesure de l'urbanisation, mesurée par la densité de population et a pour objectif de capter l'effet de la densité de l'activité économique. Enfin, d mesure le niveau initial de productivité du travail dans le secteur manufacturier (ANGERIZ et al. 2008).

Nous définissons cette spécification sous R :

```
# Spécifier le modèle à estimer
verdoorn<-p~q+u+G+d
```

La prise en compte des effets de débordements spatiaux nécessite d'estimer la spécification augmentée d'un terme autorégressif spatial (FINGLETON 2000, 2001) :

$$p_{it} = b_0 + \rho \sum_{i \neq j} w_{ij} p_{jt} + b_1 q_{it} + b_2 G_{it} + b_3 u_{it} + b_4 d_{it} + \varepsilon_{it} \tag{7.33}$$

Cette spécification est justifiée théoriquement par FINGLETON 2000 et 2001 et correspond à la spécification estimable d'un modèle inspirée de la Nouvelle Économie Géographique. À des fins d'illustration, nous considérons également une spécification alternative correspondant à un modèle autorégressif spatial dans les erreurs :

$$\begin{aligned} p_{it} &= b_0 + b_1 q_{it} + b_2 G_{it} + b_3 u_{it} + b_4 d_{it} + \varepsilon_{it} \\ \varepsilon_{it} &= \alpha_i + \lambda \sum_{i \neq j} w_{ij} \varepsilon_{jt} + v_{it} \end{aligned} \tag{7.34}$$

ou :

$$\varepsilon_{it} = \lambda \sum_{i \neq j} w_{ij} \varepsilon_{jt} + v_{it} \tag{7.35}$$

L'estimation de modèles sur données de panel avec R nécessite les packages *plm* (panel sans autocorrélation spatiale, gestion d'objets `pdata.frame` adaptée au panel) et *splm* (estimation et tests pour panels spatiaux). Il convient également de charger les packages *sp*, *maps* et *maptools* pour l'importation et la gestion des objets spatiaux.

```
# packages nécessaires
library(plm)
library(splm)
library(sp)
library(maps)
library(maptools)
```

4. L'analyse originale de FINGLETON 2000, 2001 est basée sur un modèle en coupe transversale, nous l'étendons au cas des données de panel.

L'estimation des spécifications les plus courantes se fait à l'aide des commandes `spml` et `spreml` pour le maximum de vraisemblance et `spgm` pour la méthode des moments généralisée. Celles ci ont toutes une structure relativement identique avec des options additionnelles selon les cas :

```
# Maximum de Vraisemblance :
spml(formula, data, index=NULL, listw, listw2=listw, na.action,
                 model=c("within","random","pooling"),
                 effect=c("individual","time","twoways"),
                 lag=FALSE, spatial.error=c("b","kkp","none"),
                 ...)
```

Il faut dans un premier temps définir la spécification (`formula=...`) sans indiquer les effets spatiaux (qui seront définis par des options spécifiques), indiquer le nom du `pdata.frame` (`data=...`) et le `listw` nécessaire à la création des variables spatialement décalées (`listw=...`). La nature des effets spécifiques est déterminée par l'option `model` : l'utilisateur a le choix entre `pooling` pour un modèle sur données empilées, `within` pour un modèle à effets fixes ou `random` pour un modèle à effets aléatoires. On peut également définir si les effets concernent les individus ou/et les périodes grâce à l'option `effects` qui peut être posée égale à `individual`, `time` ou `twoways`. On peut également choisir si la spécification comporte des termes spatiaux : `lag=T` pour le modèle SAR ou `lag=F` dans le cas contraire. Pour finir, on peut choisir la nature de la spécification dans le modèle à effets aléatoires : `spatial.error="b"` pour une spécification à la Baltagi, `spatial.error="kkp"` pour la spécification à la KKP (KAPOOR et al. 2007) ou `spatial.error="none"` sinon.

La commande `spreml` permet d'estimer, par le maximum de vraissemblance, plus de spécifications avec effets aléatoires (`errors=`) avec la possibilité d'envisager différentes configurations parmi lesquelles celle d'introduire de la corrélation sérielle dans le terme d'erreur. Compte tenu des calculs matriciels que cela induit, elle comporte de nombreuses options pour paramétrer l'algorithme de calcul :

```
spreml(formula, data, index = NULL, w, w2=w, lag = FALSE,
          errors = c("semsrre", "semsr", "srre", "semre",
          "re", "sr", "sem","ols", "sem2srre", "sem2re"),
          pvar = FALSE, hess = FALSE, quiet = TRUE,
          initval = c("zeros", "estimate"),
          x.tol = 1.5e-18, rel.tol = 1e-15, ...)
```

Enfin, la commande `spgm` permet d'estimer les paramètres par la méthode des moments généralisée.

```
spgm(formula, data=list(), index=NULL, listw =NULL, listw2 = NULL,
         model=c("within","random"), lag = FALSE, spatial.error=TRUE,
         moments = c("initial", "weights", "fullweights"), endog = NULL,
         instruments= NULL, lag.instruments = FALSE, verbose = FALSE,
         method = c("w2sls", "b2sls", "g2sls", "ec2sls"), control = list(),
         optim.method = "nlminb", pars = NULL)
```

Les tests de spécification reprennent en grande partie ces options. Le test d'Hausman robuste à l'hétéroscédasticité est obtenu à l'aide de la commande `sphtest`. La commande `slmtest` permet de mettre en œuvre les tests de spécification de l'autocorrélation spatiale. Les tests de spécification sur le terme d'erreur (effet aléatoire, autocorélation spatiale, autocorrélation sérielle) s'effectuent à l'aide de la commande `bsjktest`. Ces tests sont facilement interprétables car l'hypothèse alternative est toujours rappelée dans l'output.

7.4.2 Les données et la matrice de poids

Notre analyse porte sur un échantillon de 1032 régions (NUTS3) européennes localisées dans 14 États membres de l'UE15 (seule la Grèce n'est pas présente dans notre échantillon). Les données sont disponibles pour la période 1991-2008. Nous agrégeons les données annuelles par période de 3 ans afin de contrôler des variations économiques de court terme (cycles). Nous obtenons un panel de 6 périodes pour lequel nous construisons les taux de croissance de la productivité du travail (p) et de la valeur ajoutée (q) dans le secteur manufacturier. Les estimations porteront donc sur 5 périodes. La figure 7.1 représente le périmètre d'étude de notre analyse.

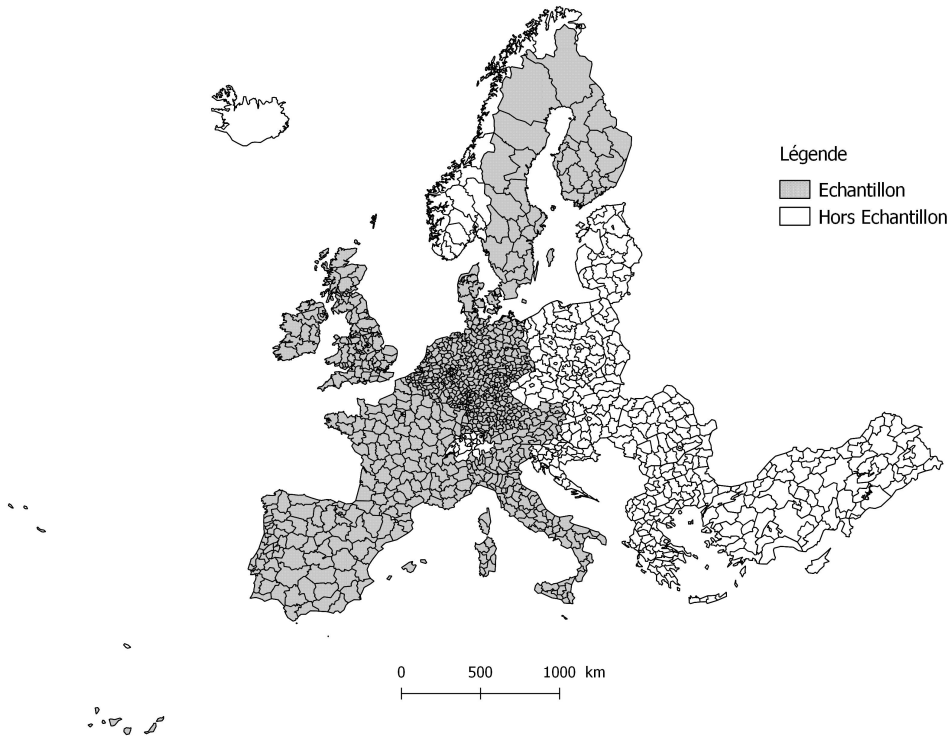


FIGURE 7.1 – Périmètre d'étude

```
# Importation des données
data_panel <- read.csv("panel_average_3_years_1991_2008.csv", sep=";")
# Importation du shapefile (Gisco) en objet "spatialpolygondataframe"
shape_nuts3<-readShapeSpatial("NUTS_RG_60M_2006")
# Sélection des NUTS3 (par niveau de NUTS)
shape_nuts3<- shape _nuts3[shape_nuts3$STAT_LEVL_== 3,]
# Sélection des NUTS3 de notre échantillon
data_panel_code<- data_panel[,"NUTS3"]
shape_nuts3<- shape _nuts3[shape_nuts3$NUTS_ID %in% data_panel_code,]
# Visualisation de l'échantillon
plot(shape_nuts3)
```

Afin de générer un tableau de statistiques descriptives en format LaTeXde la variable expliquée et des variables explicatives du modèle, il est possible d'utiliser le package `stargazer` et d'appliquer la commande *stargazer* sur la base de données comprenant les variables du modèle. Le résultat est reporté dans la table 7.2.

```
library(stargazer)
variables <-data.frame(data_panel$p,data_panel$q,data_panel$u,data_panel$G,
    data_panel$d)
stargazer(variables, title="Statistiques descriptives")
```

Statistic	N	Mean	St. Dev.	Min	Max
p	5 160	0.402	0.078	0.000	0.888
q	5 160	0.399	0.081	0.000	0.900
u	5 160	51.761	110.371	0.187	2 084.284
G	5 160	45.667	12.054	0.000	90.055
d	5 160	3.801	0.335	1.746	5.405

TABLE 7.2 – Statistiques descriptives

Concernant la matrice de poids, la présence d'îles dans l'échantillon (Madère, Canaries entre autres) nécessite l'utilisation d'une matrice de poids basée sur un autre critère que la simple contiguïté liée à la présence d'une frontière commune (voir le chapitre 2 :" Codifier la structure de voisinage"). Nous construisons une matrice des 10 plus proches voisins afin de garantir une connexion entre les régions de la Grande Bretagne et de l'Europe continentale.

```
# Création d'une matrice k plus proches voisins, k = 10
map_crd <- coordinates(shape_nuts3)
Points_nuts3 <- SpatialPoints(map_crd)
nuts3.knn_10 <- knearneigh(Points_nuts3, k=10)
K10_nb <- knn2nb(nuts3.knn_10)
wknn_10 <- nb2listw(K10_nb, style="W")
```

7.4.3 Les résultats

Pour retenir la spécification la plus appropriée, nous partons du modèle sans autocorrélation spatiale et mettons en œuvre le test d'Hausman et des tests du multiplicateur de Lagrange.

La table 7.3 présente les résultats de l'estimation d'un modèle sans autocorrélation spatiale des erreurs. La colonne (1) correspond au modèle sur données empilées alors que les colonnes (2) et (3) prennent en compte l'hétérogénéité individuelle inobservée respectivement à travers des effets fixes et des effets aléatoires. Concernant le coefficient de Verdoorn, les résultats sont similaires : avec un coefficient significatif et positif supérieur à 0.5 dans les trois cas, la présence de rendements d'échelle croissants est confirmée pour notre échantillon. Le taux de croissance de l'emploi dans le secteur manufacturier d'une région est également d'autant plus grand que cette région est urbanisée (coefficient associé à u positif et significatif dans le premier et troisième cas), d'autant plus grand que l'écart avec la région leader en début de période est important (coefficient associé à G positif et significatif dans le premier et le troisième cas) et d'autant moins important que la productivité initiale est grande, ce qui traduit un phénomène de convergence des productivités du travail dans le secteur manufacturier (coefficient associé à d négatif et significatif dans les trois cas).

```
##### Table 7.3 : estimation sans prise en compte de l'autocorrélation
    spatiale
summary(verdoorn_pooled <- plm(verdoorn, data = data_panel, model = "
    pooling"))
```

```
summary(verdoorn_fe1<- plm(verdoorn, data = data_panel,
                           model = "within", effect="individual"))
summary(verdoorn_re1<- plm(verdoorn, data = data_panel,
                           model = "random", effect="individual"))
```

		p	
Modèle :	données empilées	effets fixes (*within*)	effets aléatoires (MCG)
	(1)	(2)	(3)
q	0.692^{***}	0.604^{***}	0.701^{***}
	(0.009)	(0.010)	(0.010)
u	0.0001^{***}	-0.0002	0.0001^{***}
	(0.00001)	(0.0002)	(0.00001)
G	0.0001	0.002^{***}	0.0003^{***}
	(0.0001)	(0.0001)	(0.0001)
d	-0.008^{***}	-0.182^{***}	-0.033^{***}
	(0.003)	(0.005)	(0.003)
Constante	0.146^{***}		0.228^{***}
	(0.012)		(0.014)
Observations	5 160	5 160	5 160
$R^2 ajust$	0.523	0.587	0.552

TABLE 7.3 – Estimations sans prise en compte de l'autocorrélation spatiale
Note : * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Les résultats du test d'Hausman standard et du test d'Hausman robuste à l'autocorrélation spatiale des erreurs conduisent au rejet de l'hypothèse nulle de l'absence de corrélation entre les effets individuels et les variables explicatives. Nous optons donc dans la suite de l'analyse empirique pour un modèle à effets fixes.

```
# Test d'Hausman (plm)
print(hausman_panel<-phtest(verdoorn, data = data_panel))
## Hausman Test
## data:  verdoorn
## chisq = 1040.8, df = 4, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent

# Test d'Hausman robuste à l'autocorrélation spatiale (splm)
print(spat_hausman_ML_SEM<-sphtest(verdoorn,data=data_panel,
                        listw =wknn_10, spatial.model = "error", method="ML
                           "))
## Hausman test for spatial models
## data:  x
```

```
## chisq = 1263.8, df = 4, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent

print(spat_hausman_ML_SAR<-sphtest(verdoorn,data=data_panel,
            listw =wknn_10,spatial.model = "lag", method="ML"))
## Hausman test for spatial models
## data:  x
## chisq = 1504, df = 4, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

Les résultats des tests du multiplicateur de Lagrange dans un modèle à effets fixes nous conduisent à privilégier une spécification SEM (code des tests ci-dessous). Si les statistiques de test de prise en compte de l'autocorrélation spatiale par un SAR (Test 1) ou par un SEM (Test 2) confirment le rejet de l'hypothèse que ces deux termes (pris indépendamment) sont nuls, la lecture simultanée ne nous permet pas de conclure sur la spécification la plus appropriée pour prendre en compte l'autocorrélation spatiale (ces deux tests n'étant pas emboîtés). On peut toutefois noter que la statistique de test pour une alternative SEM est supérieure à celle correspondant à une alternative SAR. Pour conclure de façon plus crédible, on utilise des tests robustes à la présence de la spécification alternative de l'autocorrélation spatiale (Tests 3 et 4). Autrement dit, il s'agit pour le RLMlag de tester l'absence de terme autorégressif spatial lorsque le modèle contient déjà un terme autorégressif spatial dans les erreurs (RLMlag), ou inversement pour RLMerr de tester l'absence de terme autorégressif spatial dans les erreurs lorsque le modèle contient un terme autorégressif spatial. La version robuste RLMerr est fortement significative (Test 4) alors que RLMlag ne l'est pas (Test 3). Nous estimons donc un modèle à effets fixes avec un processus autorégressif spatial dans les erreurs. Dans certains cas, ces deux derniers tests robustes ne permettent pas de discriminer entre un SAR et un SEM. Plusieurs possibilités sont envisageables. La première consiste à estimer un modèle comportant ces deux termes spatiaux (SARAR). La seconde consiste à discriminer entre les deux spécifications sur les bases des statistiques des tests RLMerr et RLMlag (en prenant la spécification dont la statistique associée est la plus élevée) ou de comparer les critères d'Akaike des deux spécifications.

```
# Modèle effets fixes
# Test 1
slmtest(verdoorn, data=data_panel, listw = wknn_10, test="lml",
        model="within")
## LM test for spatial lag dependence
## data:  formula (within transformation)
## LM = 326.41, df = 1, p-value < 2.2e-16
## alternative hypothesis: spatial lag dependence

# Test 2
slmtest(verdoorn, data=data_panel, listw = wknn_10, test="lme",
        model="within")
## LM test for spatial error dependence
## data:  formula (within transformation)
## LM = 1115.5, df = 1, p-value < 2.2e-16
## alternative hypothesis: spatial error dependence

# Test 3
slmtest(verdoorn, data=data_panel, listw = wknn_10, test="rlml",
```

```
        model="within")
## Locally robust LM test for spatial lag dependence sub spatial error
## data:  formula (within transformation)
## LM = 0.0025551, df = 1, p-value = 0.9597
## alternative hypothesis: spatial lag dependence

# Test 4
slmtest(verdoorn, data=data_panel, listw = wknn_10, test="rlme",
        model="within")
## Locally robust LM test for spatial error dependence sub spatial lag
## data:  formula (within transformation)
## LM = 789.08, df = 1, p-value < 2.2e-16
## alternative hypothesis: spatial error dependence
```

		p		
Modèle :	données empilées	effets fixes (MV)		effets fixes (MMG)
		erreur Baltagi	erreur KKP	
	(1)	(2)	(3)	(4)
q	0.716***	0.650***	0.650***	0.836***
	(0.017)	(0.008)	(0.008)	(0.009)
u	0.0001***	0.0001	0.0001	0.0001
	(0.00001)	(0.0002)	(0.0002)	(0.0002)
G	-0.0004***	0.001***	0.001***	0.0003***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
d	-1.70***	-0.163***	-0.163***	-0.164***
	(0.003)	(0.0005)	(0.0005)	(0.005)
Constante	0.2***			
	(0.02)			
λ		0.566***	0.566***	0.513***
		(0.02)	(0.02)	(0.02)
Observations	5 160	5 160	5 160	5 160

TABLE 7.4 – Estimations du modèle sur données empilées et du modèle à effets fixes avec autocorrélation spatiale des erreurs
Note : * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

La table 7.4 synthétise les résultats de l'estimation du modèle avec prise en compte de l'autocorrélation spatiale sous la forme d'une autocorrélation spatiale des erreurs. Contrairement au modèle SAR, les paramètres estimés d'un SEM s'interprètent de manière classique [5]. La première colonne

5. Il n'est pas nécessaire de calculer les effets directs, indirects et totaux dans le cadre d'un SEM en raison de

correspond au modèle sur données empilées alors que les trois colonnes suivantes présentent les résultat correspondant au modèle à effets fixes avec différentes méthodes d'estimation (maximum de vraisemblance pour les colonnes (2) et (3) ; MMG pour la colonne (4)) et différentes spécifications du terme d'erreur (Baltagi pour la colonne (2) et KKP pour la colonne (3)). Dans tous les cas, le coefficient d'autocorrélation est positif et significatif. Concernant le coefficient de Verdoorn, il reste positif et significatif et d'une ampleur plus importante que précédemment. L'impact de l'urbanisation n'est plus significatif lorsque l'on introduit un effet fixe : les variations temporelles de densité de population n'affectent pas significativement le taux de croissance de de la productivité du travail. L'effet de l'urbanisation observé sur données empilées provient certainement de caractéristiques inobservables favorables à l'urbanisation (par exemple les avantages de localisation de première nature, KRUGMAN 1999).

```
##### Table 7.4 : Estimations du modèle sur données
empilées et du  modèle à effets fixes
avec autocorrélation spatiale des erreurs
# Estimation par le Maximum de Vraisemblance
summary(verdoorn_SEM_pool <- spml(verdoorn, data = data_panel,
listw = wknn_10, lag=FALSE,model="pooling"))
# SEM à effets fixes
summary(verdoorn_SEM_FE<- spml(verdoorn, data = data_panel,
listw = wknn_10, lag=FALSE,model="within", effect="individual", spatial.
    error="b"))
summary(verdoorn_SEM_FE<- spml(verdoorn, data = data_panel,
listw = wknn_10, lag=FALSE,model="within", effect="individual", spatial.
    error="kkp"))
# Estimation par la méthode des moments généralisée
summary(verdoorn_SEM_FE_GM <- spgm(verdoorn, data=data_panel,
       listw = wknn_10, model="within", moments="fullweights",
       spatial.error = TRUE))
```

7.5 Extensions

Dans cette section nous présentons certaines extensions des modèles spatiaux sur données de panel. Les méthodes présentées dans ces extensions ne sont pas implémentées dans R à l'heure actuelle.

7.5.1 Modèles dynamiques spatiaux

Les modèles considérés dans les sections précédentes sont des modèles statiques. Cependant, les interactions spatiales peuvent présenter un caractère dynamique. Ainsi, les valeurs prises pour une observation i à une période de temps t peuvent dépendre des valeurs prises par les observations voisines de i à la période précédente. Le même type de schéma peut s'appliquer pour les termes d'erreurs. Le caractère dynamique peut être pris en compte en repartant de l'équation 7.6, où nous introduisons des retards temporels sur la variable expliquée et son décalage spatial :

$$y_t = \tau y_{t-1} + \rho W_N y_t + \eta W_N y_{t-1} + x_t \beta + W_N x_t \theta + \alpha + u_t \tag{7.36}$$

l'absence d'effet multiplicateur spatial. Toutefois, nous renvoyons le lecteur à (PIRAS 2014) pour le calcul de ces effets dans un SAR en panel statique.

Ce modèle peut s'interpréter comme un modèle de Durbin spatial dynamique (DEBARSY et al. 2012 ; LEE et al. 2015). Dans ce modèle la valeur de la variable expliquée prise pour une observation i au temps t dépend de la valeur de la variable expliquée pour l'observation i à la période précédente (retard temporel), de la valeur de la variable expliquée pour les observations voisines à i à la période t (décalage spatial simultané) et enfin de la valeur de la variable expliquée pour les observations voisines de i à la période précédente $t-1$ (décalage spatial retardé). Pour ce dernier terme, on peut par exemple penser à des effets de diffusion spatiale : un choc se produisant en une zone i à une période t qui se diffuse aux zones voisines dans les périodes suivantes. On pourrait également incorporer des retards temporels sur les variables explicatives X_t ou le terme d'erreur u_t mais comme le montrent ANSELIN et al. 2008 et ELHORST 2012, les paramètres d'un tel modèle ne sont pas identifiables. Enfin, en toute généralité, ce modèle peut inclure un effet individuel, fixe ou aléatoire. DEBARSY et al. 2012 détaillent la nature des impacts (directs, indirects, totaux) dans ce modèle. Pour donner l'intuition de ces impacts, on réécrit le modèle décrit par l'équation 7.36 sous la forme suivante :

$$y_t = (I_N - \rho W_N)^{-1}(\tau y_{t-1} \eta W_N y_{t-1}) + (I_N - \rho W_N)^{-1}(x_t\beta + W_N x_t\theta) + (I_N - \rho W_N)^{-1}(\alpha + u_t) \quad (7.37)$$

La matrice des dérivées partielles de la valeur espérée de y_t par rapport à la k^{me} variable explicative de X à la période t est alors :

$$\left[\frac{\partial q\mathbb{E}(y)}{\partial x_{1k}} \quad \ldots \quad \frac{\partial q\mathbb{E}(y)}{\partial x_{nk}}\right]_t = (I_N - \rho W_N)^{-1}(\beta_k I_N + \theta_k W_N) \quad (7.38)$$

Ces dérivées partielles reflètent l'effet d'un changement affectant une variable explicative pour une observation i sur la variable expliquée de toutes les autres observations dans le court terme uniquement. Les effets de long terme sont définis par :

$$\left[\frac{\partial q\mathbb{E}(y)}{\partial x_{1k}} \quad \ldots \quad \frac{\partial q\mathbb{E}(y)}{\partial x_{nk}}\right]_t = [(1-\tau)I_N - (\rho+\eta)W_N]^{-1}(\beta_k I_N + \theta_k W_N) \quad (7.39)$$

Les effets directs sont constitués des éléments de la diagonale du terme à droite de l'équation 7.38 ou de l'équation 7.39 et les effets indirects comme la somme des lignes ou des colonnes des éléments non diagonaux de ces matrices. Ces effets sont indépendants de la période t. Il n'y a donc pas d'effet indirect de court terme si $\rho = \theta_k = 0$ et il n'y a pas d'effet indirect de long terme si $\rho = -\eta$ et si $\theta_k = 0$.

Deux grandes catégories de méthodes ont été proposées pour estimer ce modèle. D'une part, en se basant sur le principe du maximum de vraisemblance, YU et al. 2008 construisent un estimateur pour le modèle décrit par l'équation 7.36 incluant des effets fixes individuels. Cet estimateur est étendu par LEE et al. 2010a pour un modèle incluant en outre des effets fixes temporels. L'intuition est d'estimer le modèle par la méthode du maximum de vraisemblance conditionnellement à la première observation. Ils proposent également une correction lorsque le nombre d'unités spatiales et le nombre de périodes tend vers l'infini. D'autre part, LEE et al. 2010a proposent un estimateur des Moments Généralisés optimal basé sur des conditions linéaires et des conditions quadratiques. Cet estimateur est convergent, même si le nombre de périodes est petit par rapport au nombre d'observations spatiales.

Le lecteur pourra se reporter à ELHORST 2012 ou LEE et al. 2015 pour une présentation plus détaillée des modèles de panels spatiaux dynamiques.

7.5.2 Modèles multidimensionnels spatiaux

Dans certains cas, les données de panel présentent une structure multidimensionnelle plus complexe. Par exemple, dans les modèles gravitaires, des flux économiques (flux de commerce, d'IDE, etc.) entre des objets spatiaux (des pays ou des régions) sont modélisés dans des modèles de panel à trois dimensions en introduisant des effets fixes individuels, temporels, voire des effets bilatéraux d'interaction. L'introduction de l'autocorrélation spatiale dans ces modèles de type gravitaire est abordée par exemple par Arbia (2015). La structure multidimensionnelle peut également être de nature hiérarchique. Ainsi, les données régionales européennes sont disponibles à plusieurs échelles spatiales : NUTS3, NUTS2, NUTS1, les régions NUTS3 étant imbriquées dans les régions NUTS2, ces dernières étant elles-mêmes imbriquées dans les régions NUTS1. Dans le cas des modèles de panel a-spatiaux, une série d'articles des années 2000 (par exemple BALTAGI et al. 2001) modélisent cette structure hiérarchique à travers une spécification particulière des effets aléatoires. Récemment, des auteurs ont étendu cette littérature sur les modèles hiérarchiques à l'analyse des panels spatiaux (voir LE GALLO et al. 2017 pour une revue de littérature récente). Nous présentons ici la logique générale de ces modélisations.

Formellement, soit un panel multidimensionnel à 3 dimensions où la variable dépendante est observée selon trois indices : y_{ijt} avec $i = 1,2,\ldots,N$, $j = 1,2,\ldots,M_i$ et $t = 1,2,\ldots,T$. N est le nombre de groupes. M_i est le nombre d'individus dans le groupe i, de telle sorte qu'il y a $S = \sum_{i=1}^{N} M_i$ individus. T représente le nombre de périodes. En toute généralité il peut y avoir un nombre différent d'individus entre les N groupes, cependant le panel reste cylindré dans la dimension temporelle. Dans le cas d'une structure hiérarchique spatiale, on suppose que l'indice j se réfère aux individus (par exemple, les régions NUTS3) qui sont imbriqués dans N groupes (par exemple, les régions NUTS2). En supposant que l'autocorrélation spatiale se produit au niveau des individus et que les coefficients sont homogènes, on peut écrire le modèle SDM suivant :

$$y_{ijt} = \rho \sum_{g=1}^{N} \sum_{h=1}^{M_g} w_{ij,gh} y_{ght} + x_{ijt}\beta + \sum_{g=1}^{N} \sum_{h=1}^{M_g} w_{ij,gh} x_{ght}\theta + \varepsilon_{ijt}, \tag{7.40}$$

où y_{ijt} est la valeur de la variable dépendante pour l'individu j dans le groupe i à la période t. x_{ijt} est un vecteur $(1,K)$ de variables explicatives exogènes, alors que β et θ sont des vecteurs $(K,1)$ de paramètres inconnus à estimer. ε_{ijt} est le terme d'erreur avec des propriétés détaillées plus bas. La pondération spatiale $w_{ij,gh} = w_{k,l}$ est l'élément $(k = ij; l = gh)$ de la matrice de pondération spatiale W_S avec ij dénotant l'individu j dans le groupe i, et de façon similaire pour gh. Ainsi, $k,l = 1,\ldots,S$ et W_S est une matrice de pondération de dimension (S,S) avec les propriétés habituelles. ρ est le paramètre de décalage spatial. En toute généralité, on peut également spécifier une autocorrélation spatiale des erreurs, sous la forme d'un modèle autorégressif au niveau individuel :

$$\varepsilon_{ijt} = \lambda \sum_{g=1}^{N} \sum_{h=1}^{M_g} m_{ij,gh} \varepsilon_{ght} + u_{ijt}. \tag{7.41}$$

Le poids $m_{ij,gh}$ est un élément de la matrice de poids M_S. Par simplicité, on peut supposer que $M_S = W_S$. λ est le paramètre spatial à estimer. u_{ijt} est un terme aléatoire composé qui capte la structure hiérarchique des données. À cette fin, on suppose que u_{ijt} est la somme d'une composante spécifique au groupe et invariante dans le temps α_i, une composante spécifique au couple individu-groupe invariante dans le temps μ_{ij} et un terme résiduel v_{ijt} :

$$u_{ijt} = \alpha_i + \mu_{ij} + v_{ijt}, \tag{7.42}$$

avec les hypothèses suivantes : (i) $\alpha_i \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_\alpha^2\right)$, (ii) $\mu_{ij} \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_\mu^2\right)$, (iii) $v_{ijt} \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_v^2\right)$ et (iv) les trois termes sont indépendants les uns des autres. Le lecteur pourra consulter (LE GALLO et al. 2017) pour les méthodes d'estimation (maximum de vraisemblance, méthode des moments généralisés), d'inférence statistique et de prévision adaptées à ces modèles.

7.5.3 Modèles de panels à facteurs communs

L'apport majeur des données de panel réside dans la modélisation de l'hétérogénéité inobservée. Les modèles présentés précédemment se proposent de modéliser l'hétérogénéité inobservée en utilisant une transformation des variables (modèle à effets fixes) ou en posant des hypothèses sur la structure du terme d'erreur (modèle à effets aléatoires). Dans les deux cas, une restriction est faite sur la forme de l'hétérogénéité : pour chaque individu, elle est constante dans la dimension temporelle. Autrement dit, il y a une séparation totale des deux dimensions individuelle et temporelle : les effets spécifiques individuels variant entre individus mais restant constants dans le temps et les effets spécifiques temporels variant dans le temps mais constants dans la dimension individuelle. Si cette hypothèse reste crédible dans le cadre des panels courts, elle est trop restrictive pour les panels composés d'une dimension temporelle importante.

Dans certains cas, les bases de données comprennent également une dimension temporelle importante. Les modèles à facteurs commun ont été développés pour exploiter cette configuration des données. Cette nouvelle classe de modèles permet de modéliser l'effet de facteurs communs qui affectent différemment les individus, en résumant l'information présente dans les données en un nombre réduit de facteurs communs :

$$y_{it} = x_{it}\beta + \sum_{l=1}^{d} \lambda_{il} f_{lt} + \varepsilon_{it} \tag{7.43}$$

où $\sum_{l=1}^{d} \lambda_{il} f_{lt}$ correspond aux facteurs communs du modèle. Nous renvoyons à BAI et al. 2016 pour une présentation plus précise de cette classe de modèles, et nous nous intéressons à ce qui les relie aux panels spatiaux.

Par définition, les facteurs communs et panels spatiaux permettent de capter les interactions entre individus. Ils adoptent toutefois des logiques différentes. Les modèles d'économétrie spatiale reposent sur une structure donnée des interactions entre les individus d'un panel. Cette structure est généralement construite à partir d'une métrique géographique (distance entre les individus). Dans les panels à facteurs communs, la structure des interactions n'est pas contrainte *a priori* (seul le nombre de facteurs communs est contraint).

Initialement, les panels spatiaux étaient utilisés pour des panels comprenant un grand nombre d'individus (relativement à la dimension temporelle) et l'utilisation des modèles à facteurs communs employés lorsque la dimension temporelle suffisamment grande pour construire correctement les facteurs communs. Récemment, une série de travaux a mis en avant, à travers des applications, les synergies entre les deux approches (BHATTACHARJEE et al. 2011 ; ERTUR et al. 2015) et a proposé des méthodes combinant effets spatiaux et facteurs communs (PESARAN et al. 2009 ; 2011 ; SHI et al. 2017a ; 2017b). Une application récente est proposée par VEGA et al. 2016 qui étudie l'évolution des disparités de chômage entre régions néerlandaises à l'aide d'un modèle prenant en compte les dépendances spatiales et temporelles mais également la présence de facteurs communs. Leur étude met l'accent sur l'importance de prendre en compte simultanément ces trois dimensions (et non à l'aide de méthodes en plusieurs étapes) sous risque d'obtenir des résultats biaisés. L'analyse de leurs résultats suggère que la dépendance spatiale reste un élément important pour comprendre les dispersions de taux de chômage régionaux, même une fois prise en compte la dépendance temporelle et la présence de facteurs communs.

Conclusion

L'économétrie spatiale sur données de panel est aujourd'hui l'un des domaines les plus actifs de l'économétrie spatiale, tant sur le plan théorique qu'empirique. Dans ce contexte, ce chapitre a présenté les principaux modèles d'économétrie spatiale sur données de panel. Il n'a pas pour vocation d'être exhaustif sur l'ensemble des spécifications, méthodes d'estimation et d'inférence, mais il s'est concentré sur les procédures implémentables actuellement dans le logiciel R. Ces procédures concernent les modèles spatiaux de panel statiques, pour données cylindrées, avec des matrices de poids invariantes dans le temps. Des librairies ou scripts existent également pour des logiciels propriétaires comme Matlab (commandes proposés par ELHORST 2014a) et Stata (module `XSMLE`, BELOTTI et al. 2017b) et permettent de compléter les procédures proposées sous R.

Références - Chapitre 7

ANGERIZ, Alvaro, John MCCOMBIE et Mark ROBERTS (2008). « New estimates of returns to scale and spatial spillovers for EU Regional manufacturing, 1986—2002 ». *International Regional Science Review* 31.1, p. 62–87.

ANSELIN, Luc, Julie LE GALLO et Hubert JAYET (2006). « Spatial panel econometrics ». *The econometrics of panel data, fundamentals and recent developments in theory and practice*. Sous la dir. de Dordrecht KLUWER. 3e éd. T. 4. The address of the publisher : Matyas L, Sevestre P, p. 901–969.

— (2008). « Spatial panel econometrics ». *The econometrics of panel data*. Springer, p. 625–660.

BAI, Jushan et Peng WANG (2016). « Econometric analysis of large factor models ». *Annual Review of Economics* 8, p. 53–80.

BALTAGI, Badi H, Peter EGGER et Michael PFAFFERMAYR (2013). « A Generalized Spatial Panel Data Model with Random Effects ». *Econometric Reviews* 32.5, p. 650–685.

BALTAGI, Badi H et Long LIU (2008). « Testing for random effects and spatial lag dependence in panel data models ». *Statistics & Probability Letters* 78.18, p. 3304–3306.

BALTAGI, Badi H, Heun Song SEUCK et Won KOH (2003). « Testing panel data regression models with spatial error correlation ». *Journal of econometrics* 117.1, p. 123–150.

BALTAGI, Badi H, Seuck Heun SONG et Byoung Cheol JUNG (2001). « The unbalanced nested error component regression model ». *Journal of Econometrics* 101.2, p. 357–381.

BALTAGI, Badi H et al. (2007). « Testing for serial correlation, spatial autocorrelation and random effects using panel data ». *Journal of Econometrics* 140.1, p. 5–51.

BELOTTI, Federico, Gordon HUGHES, Andrea Piano MORTARI et al. (2017b). « XSMLE : Stata module for spatial panel data models estimation ». *Statistical Software Components*.

BHATTACHARJEE, Arnab et Sean HOLLY (2011). « Structural interactions in spatial panels ». *Empirical Economics* 40.1, p. 69–94.

DEBARSY, Nicolas et Cem ERTUR (2010). « Testing for spatial autocorrelation in a fixed effects panel data model ». *Regional Science and Urban Economics* 40.6, p. 453–470.

DEBARSY, Nicolas, Cem ERTUR et James P LESAGE (2012). « Interpreting dynamic space–time panel data models ». *Statistical Methodology* 9.1, p. 158–171.

ELHORST, J Paul (2003). « Specification and estimation of spatial panel data models ». *International regional science review* 26.3, p. 244–268.

— (2012). « Dynamic spatial panels : models, methods, and inferences ». *Journal of geographical systems* 14.1, p. 5–28.

— (2014a). « Matlab software for spatial panels ». *International Regional Science Review* 37.3, p. 389–405.

— (2014b). « Spatial panel data models ». *Spatial Econometrics*. Springer, p. 37–93.

ERTUR, Cem et Antonio MUSOLESI (2015). « Weak and Strong cross-sectional dependence : a panel data analysis of international technology diffusion ». *SEEDS Working Papers 1915*.

FINGLETON, Bernard (2000). « Spatial econometrics, economic geography, dynamics and equilibrium : a 'third way' ? » *Environment and planning A* 32.8, p. 1481–1498.

— (2001). « Equilibrium and economic growth : spatial econometric models and simulations ». *Journal of regional Science* 41.1, p. 117–147.

FINGLETON, Bernard et John SL MCCOMBIE (1998). « Increasing returns and economic growth : some evidence for manufacturing from the European Union regions ». *Oxford Economic Papers* 50.1, p. 89–105.

HAUSMAN, Jerry (1978). « Specification Tests in Econometrics ». *Econometrica* 46.6, p. 1251–1271.

HSIAO, Cheng (2014). *Analysis of panel data*. 54. Cambridge university press.

KAPOOR, Mudit, Harry H KELEJIAN et Ingmar R PRUCHA (2007). « Panel data models with spatially correlated error components ». *Journal of Econometrics* 140.1, p. 97–130.

KELEJIAN, Harry H et Ingmar PRUCHA (1998). « A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances ». *Journal of Real Estate Finance and Economics* 17, p. 99–121.

— (1999). « A generalized moments estimator for the autoregressive parameter in a spatial model ». *International Economic Review* 40.2, p. 509–533.

KRUGMAN, Paul (1999). « The role of geography in development ». *International regional science review* 22.2, p. 142–161.

LE GALLO, Julie et Alain PIROTTE (2017). « Models for Spatial Panels ».

LEE, Lung-fei et Jihai YU (2010a). « A spatial dynamic panel data model with both time and individual fixed effects ». *Econometric Theory* 26.2, p. 564–597.

— (2010b). « Some recent developments in spatial panel data models ». *Regional Science and Urban Economics* 40.5, p. 255–271.

— (2015). « Spatial panel data models ».

LESAGE, James et Robert K PACE (2009). *Introduction to spatial econometrics*. Chapman et Hall/CRC.

MANSKI, Charles F (1993). « Identification of Endogenous Social Effects : The Reflection Problem ». *Review of Economic Studies* 60.3, p. 531–542.

MILLO, Giovanni (2014). « Maximum likelihood estimation of spatially and serially correlated panels with random effects ». *Computational Statistics and Data Analysis* 71, p. 914–933.

MILLO, Giovanni et Gianfranco PIRAS (2012). « splm : Spatial panel data models in R ». *Journal of Statistical Software* 47.1, p. 1–38.

MUTL, Jan et Michael PFAFFERMAYR (2011). « The Hausman test in a Cliff and Ord panel model ». *The Econometrics Journal* 14.1, p. 48–76.

PESARAN, M Hashem et Elisa TOSETTI (2009). « Large panels with spatial correlations and common factors ». *Journal of Econometrics* 161.2, p. 182–202.

— (2011). « Large panels with common factors and spatial correlation ». *Journal of Econometrics* 161.2, p. 182–202.

PIRAS, Gianfranco (2014). « Impact estimates for static spatial panel data models in R ». *Letters in Spatial and Resource Sciences* 7.3, p. 213–223.

SHI, Wei et Lung-fei LEE (2017a). « Spatial dynamic panel data models with interactive fixed effects ». *Journal of Econometrics* 197.2, p. 323–347.

— (2017b). « A spatial panel data model with time varying endogenous weights matrices and common factors ». *Regional Science and Urban Economics*.

VEGA, Solmaria Halleck et J Paul ELHORST (2016). « A regional unemployment model simultaneously accounting for serial dynamics, spatial dependence and common factors ». *Regional Science and Urban Economics* 60, p. 85–95.

VERDOORN, JP (1949). « On the factors determining the growth of labor productivity ». *Italian economic papers* 2, p. 59–68.

YU, Jihai, Robert DE JONG et Lung-fei LEE (2008). « Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and T are large ». *Journal of Econometrics* 146.1, p. 118–134.

8. Lissage spatial

LAURE GENEBES, AURIANE RENAUD ET FRANÇOIS SÉMÉCURBE
Insee

Résumé

Le lissage spatial est l'une des méthodes essentielles pour analyser les données et l'organisation spatiale. L'idée est de filtrer l'information pour révéler des structures spatiales sous-jacentes.

Du point de vue conceptuel, le lissage spatial est une méthode d'estimation non paramétrique de la fonction d'intensité d'un processus ponctuel à valeurs dans $\mathbb{R}^2$, à partir uniquement d'une de ses réalisations (que l'on observe). La fonction d'intensité théorique en un point x est obtenue en calculant la moyenne des points observés par unité de surface sur des voisinages contenant x, ces voisinages étant de plus en plus petits.

Mais, en pratique, on dispose d'une seule réalisation (observée), et cette approche par passage à la limite n'a plus de sens. Les méthodes non paramétriques à noyau contournent cette limitation, en ne proposant pas directement une estimation de la fonction d'intensité mais une estimation lissée de celle-ci. Au prix de cette approximation, lorsque le paramètre de bande passante est bien choisi, les estimations obtenues sont statistiquement robustes et géographiquement pertinentes, et permettent de déceler si la fonction d'intensité est constante ou variable dans l'espace.

On s'inspire des outils de l'analyse spatiale pour produire des analyses géographiques pertinentes. L'idée est d'obtenir une représentation cartographique simplifiée et lisible, en s'affranchissant de l'arbitraire des découpages territoriaux, et en limitant en partie le "Modifiable Area Units Problem". Dans ce cas, la bande passante s'apparente à un paramètre de généralisation géographique qui conserve ou supprime en fonction des exigences de l'analyse les détails des phénomènes géographiques observés. En pratique, il est possible de lisser des données pondérées d'après BRUNSDON et al. 2002 : chaque point de l'espace est affecté d'une valeur numérique. Plusieurs types de lissage peuvent être menés, notamment un lissage "classique" reposant sur des

calculs locaux de moyennes, ou un lissage "quantile" utilisant des calculs locaux de quantiles (médiane, décile), voir BRUNSDON et al. 2002. De plus, des opérations sur les valeurs lissées permettent notamment de calculer des ratios "lissés", tels que la part d'une sous-population dans l'ensemble de la population.

La mise en œuvre d'un lissage est désormais assez aisée, en particulier avec le logiciel R dont plusieurs packages comportent des fonctions permettant de réaliser des lissages.

8.1 Lissage spatial

La fonction d'intensité théorique en un point x est obtenue en calculant la moyenne des points observés par unité de surface sur des voisinages contenant x (voir chapitre 4 : "Les configurations de points") de plus en plus petits. Le lissage spatial est une méthode d'estimation non paramétrique de la fonction d'intensité d'un processus ponctuel à valeurs dans $\mathbb{R}^2$ à partir uniquement d'une de ses réalisations. Pour obtenir la fonction d'intensité théorique à partir de la connaissance d'une seule réalisation, on n'estime pas directement la fonction d'intensité mais une fonction de la fonction d'intensité.

D'un point de vue pratique, le lissage spatial est une modélisation locale qui repose sur le choix de paramètres.
Le noyau décrit la façon dont le voisinage est appréhendé.
La bande passante est le paramètre fondamental de l'analyse. Elle quantifie la «taille» du voisinage. Ce paramètre résulte d'un arbitrage biais-variance entre la précision spatiale de l'analyse et sa qualité statistique.
Le traitement des effets de bord explicite la façon dont les frontières géographiques et les limites du territoire d'observation sont prises en compte dans l'analyse.

Par ailleurs, on peut définir un ensemble de coordonnées géographiques pour lesquelles les valeurs lissées seront estimées (éventuellement différent de l'ensemble des coordonnées géographiques des données d'origine). La plupart des applications faites par l'Insee lissent les données sur une grille de carreaux (la nouvelle coordonnée étant le centre du carreau).

Dans ce chapitre, nous aborderons dans un premier temps les fondements et le formalisme du lissage spatial puis ses mises en œuvre.

8.1.1 Origine et formalisme du lissage spatial

Historiquement, la première méthode non paramétrique d'estimation de l'intensité repose sur la construction d'intensité territoriale. Elle consiste à calculer, pour chaque unité territoriale, l'intensité de points observée par unité de surface. Dans ce cas, l'intensité est également appelée densité. Au sein de chacune de ces unités territoriales, l'intensité estimée est constante. Par exemple, lorsqu'on calcule la densité d'une région, celle-ci est considérée identique sur tout le territoire.

L'intérêt pratique de l'intensité repose sur la possibilité de représenter les densités territoriales sous la forme de cartes choroplèthes dont les premières réalisations remontent aux travaux du Baron Pierre Charles Dupin, voir PALSKY 1991. Les géographes et statisticiens utilisèrent ensuite cette méthode pour représenter la répartition de la population dans les découpages administratifs. D'un point de vue technique, les cartes de densité généralisent les histogrammes des analyses monodimensionnelles aux espaces géographiques de dimension deux. Un exemple de carte de densité est donné sur la figure 8.1.

Au 20^e siècle, les géographes et les statisticiens se sont mis progressivement à questionner la pertinence statistique et géographique de ce type d'approche. Openshaw a théorisé ses limites sous le nom de *Modifiable Area Units Problem* (MAUP) . Le MAUP (voir figure 8.2) se décompose en

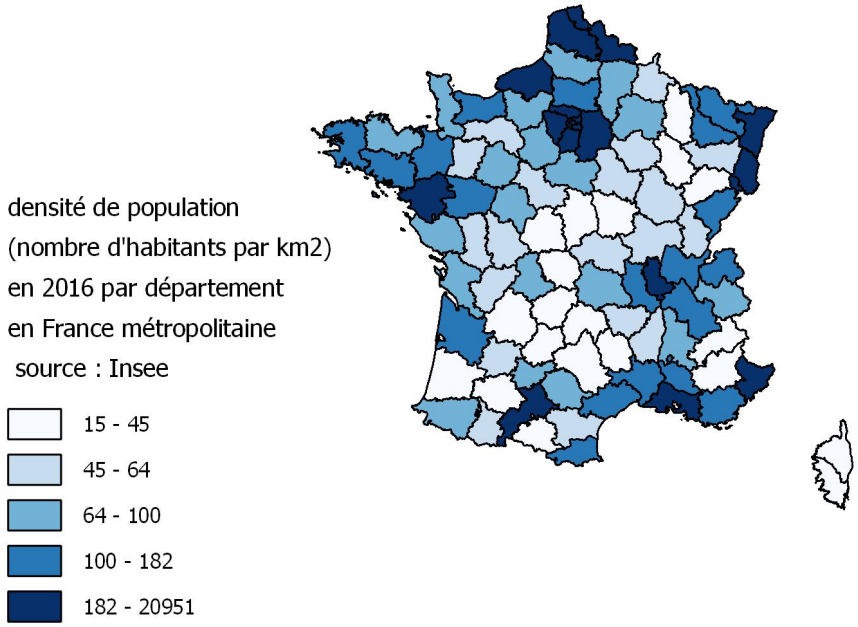


FIGURE 8.1 – Exemple de carte de densité
Source : *Insee*

deux sous-problèmes interdépendants : l'effet d'échelle (*scale effect*) et l'effet de zonage (*zoning effect*). L'effet d'échelle décrit la dépendance du phénomène observé à la taille moyenne des unités spatiales. Plus cette taille est importante et plus les spécificités locales sont réduites et plus les analyses laissent apparaître les structures globales. Au contraire, des petites tailles conservent les spécificités locales et les détaillent, mais en contrepartie les analyses sont sensibles aux bruits statistiques, à la qualité et à la précision des données. L'effet de zonage explicite la dépendance des phénomènes observés à la forme des unités spatiales. La notion de forme inclut la morphologie des unités spatiales mais également leur position dans l'espace. Ainsi, si l'on déplace uniformément le contour des unités spatiales, le phénomène observé est susceptible d'être profondément modifié.

Le lissage spatial hérite de ces réflexions et a pour objectif de s'affranchir de l'arbitraire des découpages territoriaux. Si le lissage spatial trouve une définition rigoureuse dans le cadre de l'analyse spatiale, on détecte des méthodes apparentées en géographie et en statistique dès la fin du 19^{e} siècle avec les travaux de Louis-Leger Gauthier et de Victor Turquan. Cette proximité, voire intrication, entre l'approche des statisticiens spatiaux et des géographes justifie que ce chapitre s'intéresse au lissage à la fois sous un angle d'analyse spatiale pure et d'analyse géographique plus opérationnelle.

En pratique, la difficulté à laquelle est confronté le statisticien est celle d'observer une seule réalisation. Concrètement, pour contourner cette difficulté d'estimer une fonction d'intensité à partir d' une seule réalisation, le lissage spatial n'estime pas directement celle-ci mais une version

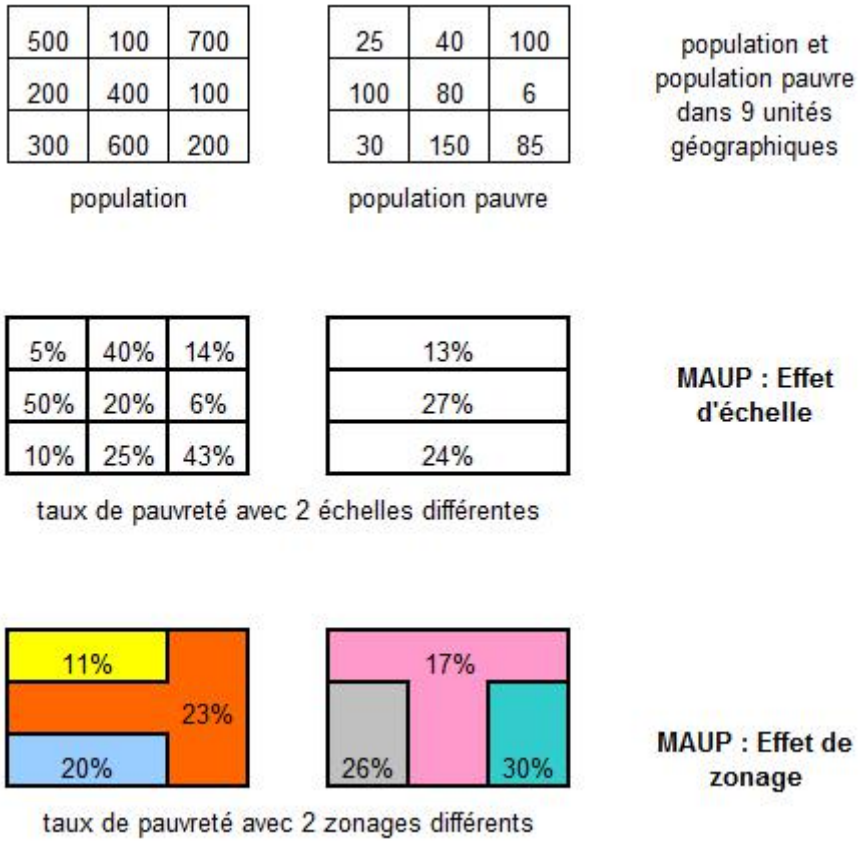


FIGURE 8.2 – Schématisation du MAUP : effet d'échelle et effet de zonage

lissée obtenue par convolution avec un noyau K_h :

$$(K_h * \lambda)(x) = \int_{\mathbb{R}^2} \lambda(t) K(x-t) dt \tag{8.1}$$

avec $K_h(u) = \frac{1}{h^2} K(\frac{u}{h})$

et K une fonction symétrique de $\mathbb{R}^2$ dans $\mathbb{R}$ positive et d'intégrale 1

R Une métaphore simple pour comprendre l'opération de convolution consiste à imaginer que λ représente la répartition de la densité de terriers de lapins dans l'espace. Pour chaque terrier est associé un unique lapin. Chaque lapin, pour subvenir à ses besoins, se déplace à proximité de son terrier de telle sorte que sa probabilité de se retrouver à une position t par rapport à son terrier est $K_h(t)$. La convolution $(K_h * \lambda)(x)$ représente dans ce cas la densité locale de lapins en x. Si h est petit, les lapins se concentrent autour de leur terrier et la fonction d'intensité des lapin diffère peu de celle des terriers. Au contraire si h est important, les lapins ont tendance à se mélanger dans l'espace et la fonction d'intensité des lapins est «floutée» par rapport à celle des terriers.

Pour obtenir un estimateur de $(K_h * \lambda)(x)$ à partir d'un ensemble de points $\{x_i\}$ issu d'une réalisation d'un processus ponctuel, une idée simple consiste à substituer l'intégrale sur $\mathbb{R}^2$ par une somme sur les points observés dans l'équation (8.1).

Définition 8.1.1 — Lissage spatial. Soit K_h un noyau de bande passante h et x un point de $\mathbb{R}^2$, l'intensité lissée estimée en x est définie par :

$$\widehat{\lambda_h}(x) = \sum_i K_h(x - x_i) \tag{8.2}$$

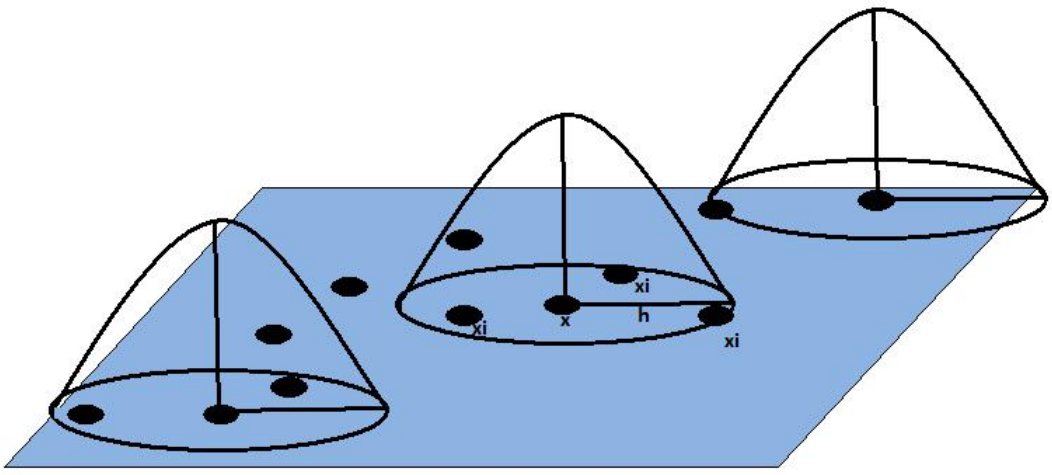


FIGURE 8.3 – Schématisation du lissage spatial avec un noyau
Note : En chaque point d'estimation figure une fonction noyau. La valeur de cette fonction est la plus élevée au niveau du point, et décroît au fur et à mesure qu'on s'en éloigne.

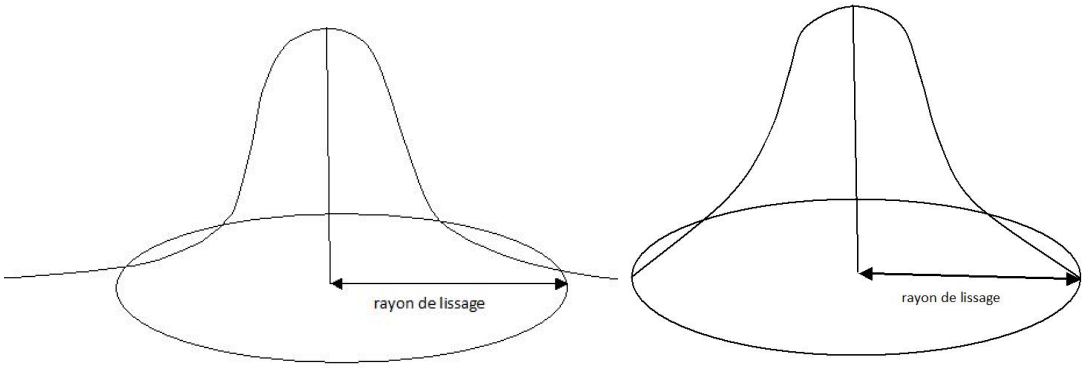
K_h dans cette formule joue un rôle analogue à une unité territoriale centrée sur chaque point de l'espace $\mathbb{R}^2$ de taille h. Contrairement aux analyses reposant sur un découpage géographique, l'estimateur de l'intensité lissée contrôle l'effet de zonage du MAUP, le choix du noyau impacte peu les résultats du lissage. En revanche, l'arbitraire de l'effet d'échelle est conservé au travers

du choix de la bande passante. Différents noyaux ont été proposés dans la littérature. Le lecteur trouvera ci-dessous les noyaux les plus fréquemment utilisés.

Définition 8.1.2 — Noyaux usuels. x est un point de $\mathbb{R}^2$. K^N et K^B sont respectivement appelés noyau gaussien et noyau quadratique :

$$K_h^N(x) = \frac{1}{2\pi}\, e^{-||\frac{x}{h}||^2} \tag{8.3}$$

$$K_h^B(x) = \frac{9}{16}\ 1_{||x||<h}\ (1-||\frac{x}{h}||^2)^2 \tag{8.4}$$



(a) Noyau gaussien K^N (b) Noyau quadratique K^B

FIGURE 8.4 – Le noyau quadratique K^B donne un poids plus élevé aux points les plus proches qu'aux points éloignés. Il s'annule au-delà du rayon de lissage. Au contraire, le noyau gaussien K^N prend en compte l'ensemble des points de la zone d'étude.

8.1.2 Traitement des effets de bord

Par rapport à l'estimation par noyau des densités de probabilités, les méthodes de lissage spatial sont impactées par un problème supplémentaire lié à la prise en compte des effets de bord. Dans le cas de l'estimation de densité, l'estimation est réalisée sur $\mathbb{R}^n$. Pour le lissage spatial, en général les points observés sont contenus dans une fenêtre d'analyse W, soit, concrètement, un polygone.

La nature des frontières de la fenêtre peut être de deux sortes. Premièrement, la fenêtre peut résulter du protocole de collecte de l'information. Par exemple, lors d'une fouille archéologique, seule une zone restreinte est fouillée pour des raisons de coûts et d'opportunités. Dans ce cas, les frontières ne sont pas inhérentes au processus observé, et, sans information complémentaire il est raisonnable de postuler la continuité de l'intensité entre l'intérieur et l'extérieur de la fenêtre. *A contrario*, la fenêtre peut être induite par des configurations géographiques qui ont un impact sur le processus sous-jacent générateur de l'ensemble de points observés. En géographie, les fleuves, les reliefs, les côtes maritimes sont autant de frontières qui contraignent l'implantation des activités humaines. À l'extérieur d'une frontière de ce type, l'intensité du phénomène observé est nulle.

La formule (8.2) est la formule d'estimation sans traitement des effets de bord. Elle revient à ignorer la fenêtre d'analyse.

Le traitement de l'effet de bord a pour objectif de prendre en compte, dans l'estimation de l'intensité, l'impact de la frontière. Différentes solutions ont été proposées. Elles se distinguent par leur façon d'appréhender l'extérieur de la zone d'observation et leur rapidité d'exécution (Baddeley, voir BADDELEY et al. 2015a).

Définition 8.1.3 — Traitement des effets de bords. x est un point de $\mathbb{R}^2$, les estimations uniforme et de Diggle (voir DIGGLE 2013) sont obtenues par les formules suivantes :

$$\text{correction uniforme : } \widehat{\lambda_h^U}(x) = \frac{1}{e_{h(x)}} \sum_i K_h(x - x_i) \tag{8.5}$$

$$\text{correction Diggle : } \widehat{\lambda_h^D}(x) = \sum_i \frac{1}{e_{h(x_i)}} K_h(x - x_i) \tag{8.6}$$

où $e_h(u) = \int_W K_h(u - v)dv$

Lorsque la fenêtre d'analyse est indépendante du processus sous-jacent, l'estimation uniforme assure la continuité de l'intensité entre l'intérieur et l'extérieur de la fenêtre. En revanche, si l'intensité en dehors de la fenêtre est jugée nulle, il est plus opportun d'utiliser l'estimation de Diggle, voir DIGGLE 2013, qui est conservative. Dans ce cas, l'intégrale de l'intensité estimée dans la fenêtre d'analyse correspond exactement aux nombres de points observés. D'un point de vue algorithmique, l'estimation de Diggle est sensiblement plus consommatrice en temps de calcul que l'estimation uniforme.

R De façon imagée, le terme $e_{h(x)}$ peut s'interpréter comme une probabilité d'intersection de deux ensembles. Supposons, toujours en prenant notre métaphore des lapins, que l'emprise spatiale de la fenêtre correspond à un enclos. $K_h(x-u)$ décrit approximativement le territoire d'exploration d'un lapin autour d'un terrier situé en x si le lapin ne rencontre pas d'obstacle. $e_h(x) = \int_W K_h(u-x)du$ est la part du territoire exploré par le lapin contenue dans l'emprise spatiale de l'enclos. $e_{h(x)}$ est strictement inférieure à 1 si le point x est à proximité immédiate de la frontière de l'enclos. En revanche, si la zone d'exploration "naturelle" des lapins du terrier est entièrement contenue dans l'enclos, $e_{h(x)}$ est égale à 1.

Dans la formule (8.5) le terme $e_{h(x)}$ est appliqué globalement à l'estimation de densité. À proximité de la frontière de l'enclos, ce terme permet de redresser l'estimation de l'intensité. Plus le point est proche de la frontière et plus $e_{h(x)}$ est faible et la compensation sera forte. La correction uniforme considère qu'à la frontière de la fenêtre, la répartition des terriers est quasi-homogène entre l'intérieur et l'extérieur de la fenêtre. Intuitivement, cette correction revient à postuler que l'enclos n'a aucun effet sur la mobilité des lapins qui la franchisse sans s'en apercevoir. Plus précisément, on considère que les lapins dont les terriers sont situés à l'extérieur de l'enclos participent également au calcul de l'intensité à l'intérieur de l'enclos.

L'estimation de Diggle, formule (8.6), suppose que la fenêtre fait partie intégrante des propriétés du processus sous-jacent. Autrement dit, l'enclos représente une frontière infranchissable pour les lapins et tous les lapins sont contenus dans l'enclos. En divisant $K_h(x-x_i)$ par le terme $e_{h(x_i)}$, on s'assure que le lapin du terrier i a une probabilité égale à 1 de rester dans l'empreinte spatiale de l'enclos.

8.1.3 Choix de la bande passante

Le choix de la bande passante conditionne l'aspect plus ou moins «lissé» de l'estimation de la fonction d'intensité. En analyse spatiale, la bande passante résulte d'un compromis biais-variance. Le biais est induit par le fait que l'estimateur de la fonction d'intensité n'estime pas directement la fonction d'intensité mais une version lissée de celle-ci. Plus la bande passante est importante et plus le biais est important. La variance décroît au contraire en fonction de la bande passante. Plus la bande passante est importante, et plus le nombre de points participant au calcul des estimations locales augmente, ce qui tend à réduire la variance d'estimation.

Plusieurs méthodes sont disponibles pour proposer automatiquement une bande passante qui minimise un critère d'erreur. Ne disposant évidemment pas de la fonction d'intensité recherchée, une partie de ces méthodes repose sur des méthodes de validation croisée. Elles se servent de la distribution de points observée, et supposent qu'elle suit une distribution de Poisson pour

estimer une bande passante optimale. Dans la section 8.3, des exemples exploitant les fonctions de validation croisée du package *spatstat* de R seront proposés. Ces exemples mettent en valeur la grande variabilité des bandes passantes proposées en fonction des critères d'erreurs choisis. Par ailleurs, l'existence d'une unique bande passante pertinente pour toute l'étendue de la zone étudiée est une hypothèse forte. Plusieurs méthodes de lissages adaptatifs ont été proposées pour dépasser cette limite. Le lecteur pourra lire avec intérêt l'ouvrage de Baddeley (BADDELEY et al. 2015a) sur cette thématique qui exploite le package *spatstat* de R.

Finalement, en soi, aucune bande passante n'est optimale : toutes sont susceptibles d'apporter une représentation du monde pertinente conformément au MAUP. Certains géographes conseillent d'adopter une démarche multi-échelle pour appréhender la pluralité des aspects spatiaux d'un même phénomène.

8.2 Lissage géographique

Le lissage géographique s'inspire de l'estimation d'intensité présentée précédemment. Il n'a pas vocation à calculer des intensités, mais à obtenir des représentations cartographiques simplifiées. Le principe de cette utilisation en géographie est de représenter non pas la valeur observée en un point, mais une moyenne pondérée des valeurs observées au voisinage de ce point dans un rayon prédéfini.

R Le lissage peut être interprété comme un outil pouvant assurer une forme de **confidentialité**. Il permet de représenter de manière agrégée des données initialement ponctuelles et confidentielles. Il faut néanmoins rester vigilant sur le nombre de points utilisés pour produire l'estimation lissée.

8.2.1 Lissage de données pondérées

On se place dans le cas où chaque point x_i est affecté d'une valeur numérique w_i. Par exemple, x_i peut représenter un logement et w_i le nombre d'habitants de ce logement. Il suffit (voir BRUNSDON et al. 2002) d'utiliser une version pondérée des estimateurs à noyaux décrits précédemment. Dans la formule (8.2), on multiplie par le poids w_i la contribution d'un point à l'estimateur d'intensité.

Définition 8.2.1 — Estimateurs à noyaux pondérés. Soit K_h un noyau de bande passante h et x_i un point de $\mathbb{R}^2$ affecté d'une pondération w_i, l'intensité lissée estimée en x est définie par :

$$\widehat{\lambda_h}(x) = \sum_i w_i K_h(x - x_i) \tag{8.7}$$

Alors que le choix du noyau K_h a peu d'influence sur les résultats du lissage (voir figure 8.5), le choix de la bande passante h est primordial, bien qu'assez arbitraire.

Comme cela a été souligné plus haut, cette bande passante se comporte comme un paramètre de lissage, contrôlant l'équilibre entre biais et variance. Un rayon élevé conduit à une densité très lissée, avec un biais élevé. Un petit rayon génère une densité peu lissée avec une forte variance. Il est généralement déterminé par l'utilisateur de faire un compromis, en fonction du niveau d'agrégation souhaité. Il est conseillé de tester plusieurs valeurs de bande passante, permettant de révéler des variations locales à différentes échelles. Les cartes de la figure 8.6 sont des exemples de cartes lissées pour Paris et sa banlieue, avec trois rayons de lissage différents.

L'intérêt de l'estimation est de s'intéresser non pas aux points et leur répartition, mais à leur environnement. La bande passante permet ainsi de définir cet environnement.

R

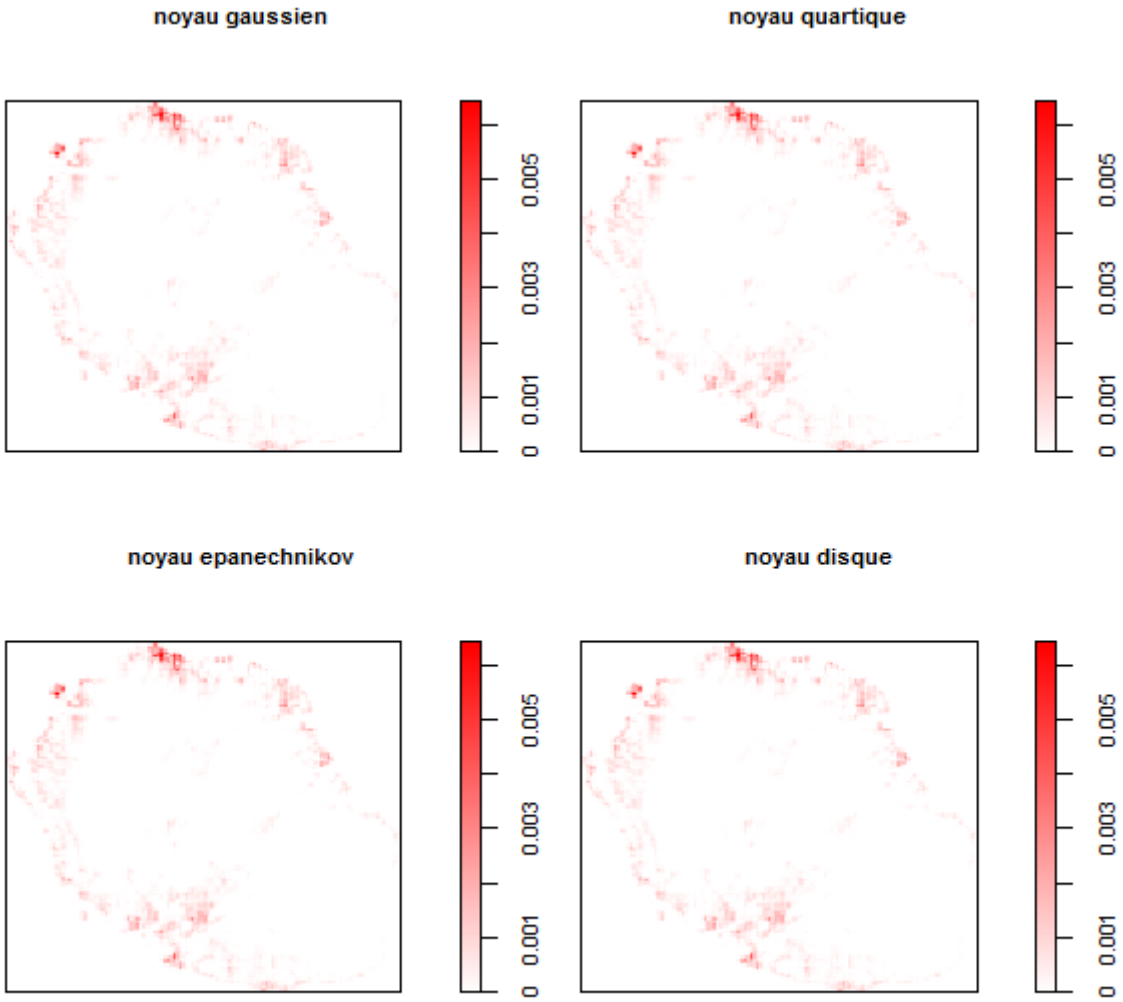
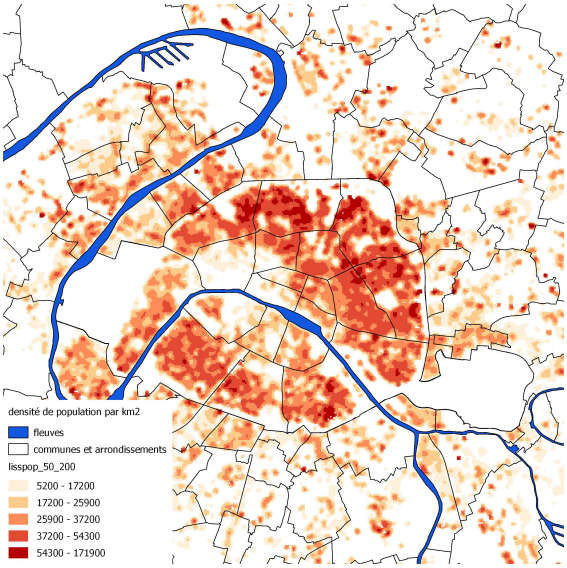
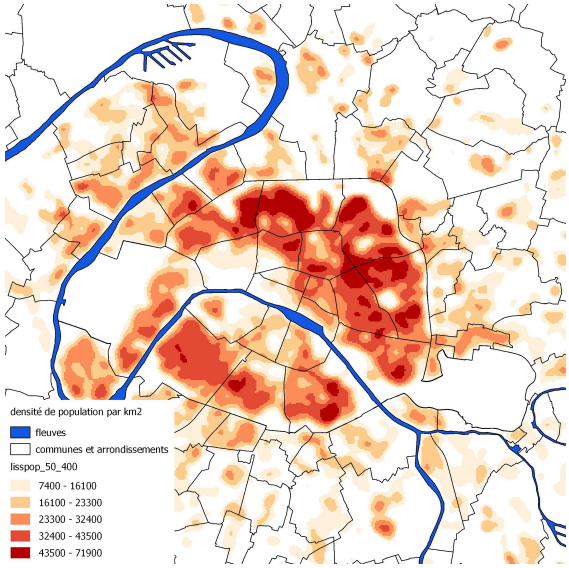


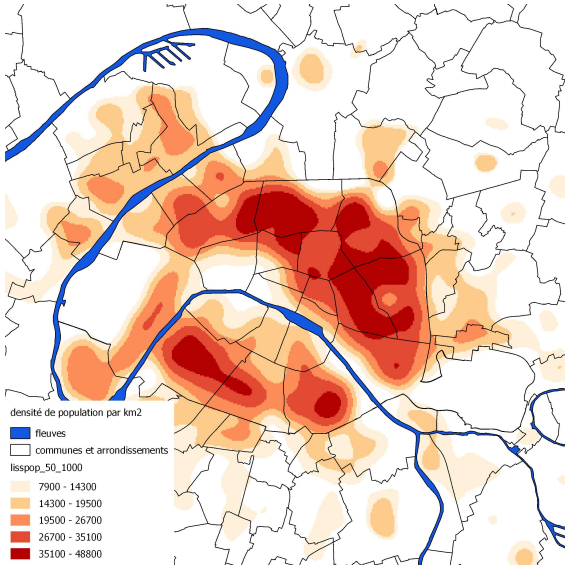
FIGURE 8.5 – Comparaison de résultats obtenus avec quatre noyaux différents à partir de la fonction `density.ppp` du package *spatstat*
Source : *Insee, Revenus Fiscaux Localises (RFL) au 31 décembre 2010 et Taxe d'habitation (TH) au 1er janvier 2011*
Champ : Ile de la Réunion
Note : La variable représentée est le nombre lissé de ménages



(a) Rayon de 200 mètres



(b) Rayon de 400 mètres



(c) Rayon de 1000 mètres

FIGURE 8.6 – Trois rayons de lissage différents pour la densité de population à Paris et sa banlieue : 200 mètres, 400 mètres, 1000 mètres
Source : *Insee-DGFiP-Cnaf-Cnav-CCMSA, Fichier localisé social et fiscal 2012*
Note : Les carreaux représentés contiennent plus de 11 ménages

Plusieurs algorithmes existent pour déterminer un rayon de lissage dit «optimal». Ces tests peuvent donner des résultats variables et parfois très éloignés (voir mise en œuvre) : il est conseillé de ne les utiliser qu'à titre indicatif, il revient à l'utilisateur de choisir le rayon de lissage compte tenu de son expérience des données, et de la problématique.

Des opérations peuvent être effectuées sur les variables lissées, notamment des ratios. La justification théorique se trouve pp. 34 à 37 du document de travail de J.M. Floch (FLOCH 2012). De manière pratique, si l'on souhaite obtenir la valeur lissée du ratio de deux variables, il est primordial de calculer séparément les valeurs lissées du numérateur et du dénominateur, et ensuite de calculer le rapport entre la valeur lissée du numérateur, et la valeur lissée du dénominateur. Ne pas calculer directement une valeur lissée d'un ratio : la carte serait déformée, car on donnerait à tort la même importance aux différents territoires, pourtant inégalement peuplés.

8.2.2 Application utilisant une régression non paramétrique

On s'intéresse au **calcul d'un revenu moyen** par personne. On dispose de deux variables : le revenu, et le nombre de personnes. Le revenu moyen est égal à la somme de l'ensemble des revenus, divisée par la somme du nombre de personnes. On lisse séparément les revenus, et le nombre de personnes. On calcule ensuite le rapport.

On obtient les cartes de la figure 8.7 pour Paris et les communes environnantes faisant partie de la "petite couronne" (*i.e.* les trois départements limitrophes de Paris) :

La carte 8.7a du niveau de vie total des ménages n'a pas beaucoup de sens, il est nécessaire de rapporter ce niveau de vie total par carreau à la population de chaque carreau.
Sur la carte 8.7b du nombre de personnes : la population est très dense au sein de la commune de Paris, essentiellement au Nord-Est de la Seine, et dans une moindre mesure à l'extrême Sud.
Sur la carte 8.7c, le niveau de vie moyen par personne est très élevé en plein cœur de Paris, essentiellement à l'Ouest.

R Ici, on n'est plus dans le cadre théorique. Ce calcul s'inspire d'outils apparentés à une régression non paramétrique. Approximativement, c'est comme si l'on faisait une régression géographique pondérée, qui se limiterait à une seule variable : la constante (voir chapitre 9 : "Régression géographiquement pondérée").

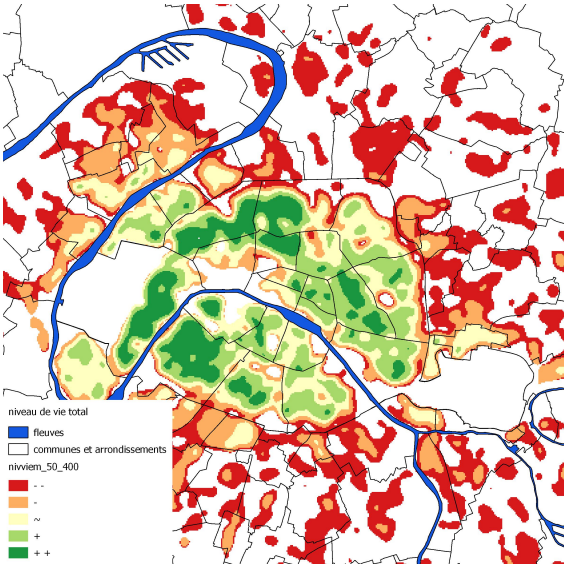
8.2.3 Application utilisant une estimation de densité conditionnelle non paramétrique

On s'intéresse à la **part des ménages pauvres** dans l'ensemble des ménages. On calcule la valeur lissée du nombre de ménages pauvres d'un territoire, et on calcule la valeur lissée du nombre total de ménages sur le territoire. On calcule ensuite le rapport.

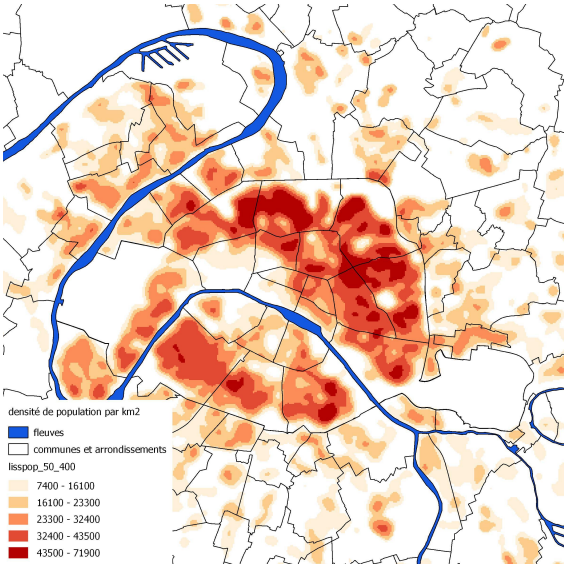
D'après la carte de la figure 8.8a, les zones les plus peuplées sont situées au cœur de Paris : à la fois le quart Nord-Est, et le quart Sud-Ouest. Dans la figure 8.8b, les ménages pauvres sont nombreux à Paris, plutôt dans le quart Nord-Est. Dans la figure 8.8c, la part des ménages pauvres dans l'ensemble des ménages apporte une information complémentaire. Sur cette carte, sont mises en exergue des zones moins densément peuplées, mais pour lesquelles la part de ménages vivant en dessous du seuil de pauvreté est forte. Il s'agit de communes situées au nord de Paris.

Ainsi, selon la carte produite, les messages obtenus peuvent être différents. Lorsque l'on analyse des taux, il est indispensable d'analyser également la répartition des simples effectifs (densités de population par exemple), pour vérifier la robustesse des taux calculés, et leur représentativité.

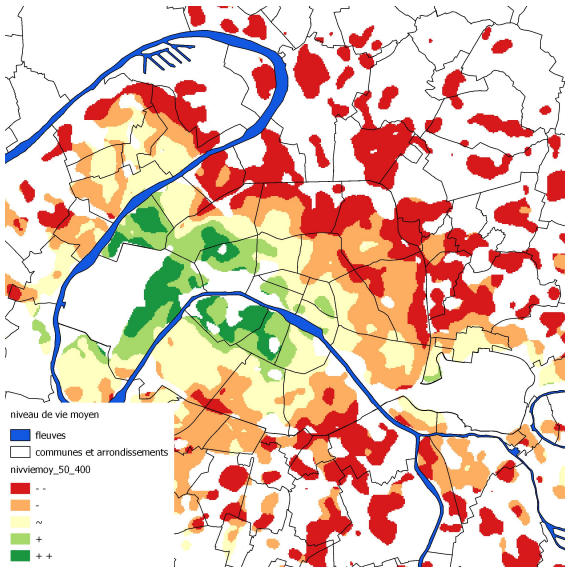
R Ce calcul s'apparente à un calcul de probabilité conditionnelle. On obtient une carte représentant des taux de pauvreté au niveau local, ce qui est proche de l'idée d'obtenir la probabilité qu'un ménage soit pauvre sachant qu'il habite à un endroit donné.



(a) Total du niveau de vie des ménages (en euros)

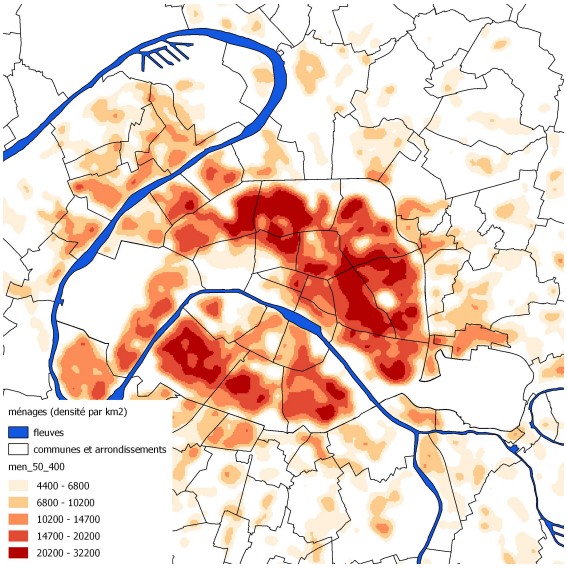


(b) Population des ménages (densité par km2)

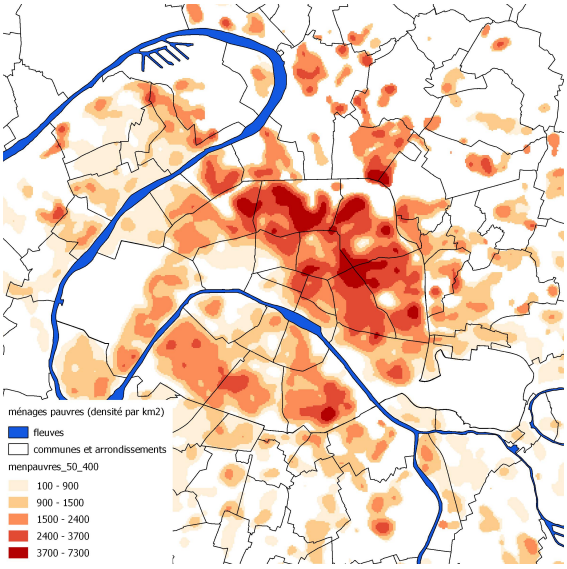


(c) Niveau de vie moyen

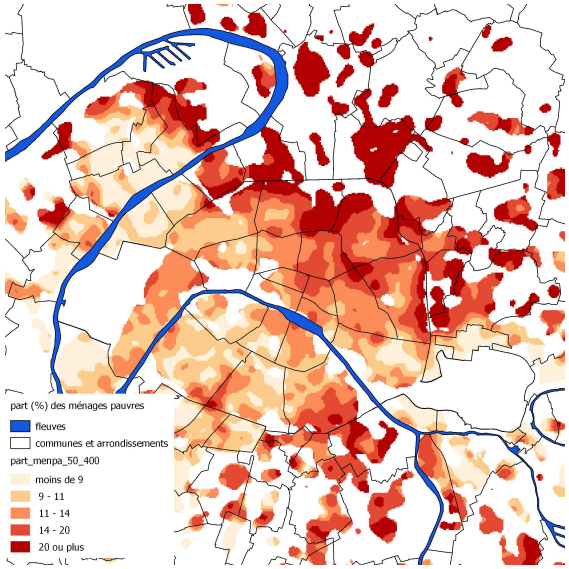
FIGURE 8.7 – Calcul d'un niveau de vie moyen lissé
Source : *Insee-DGFiP-Cnaf-Cnav-CCMSA, Fichier localisé social et fiscal 2012*
Note : Les carreaux représentés contiennent plus de 11 ménages. Pour les cartes représentant des niveaux de revenus, les marqueurs "++", "+", "∼", "-" et "- -" correspondent à des valeurs respectivement très élevées, élevées, moyennes, basses ou très basses pour l'indicateur considéré. Ils ont été utilisés pour des questions de non-profilage de la population



(a) Ménages (densité par km2)



(b) Ménages pauvres (densité par km2)



(c) Part des ménages pauvres dans l'ensemble des ménages

FIGURE 8.8 – Calcul de la part lissée des ménages pauvres
Source : *Insee-DGFiP-Cnaf-Cnav-CCMSA, Fichier localisé social et fiscal 2012*
Note : Les carreaux représentés contiennent plus de 11 ménages

Attention ! En théorie, il est toujours possible de calculer le rapport de deux variables lissées. En pratique, il est nécessaire de faire attention aux petits effectifs. Dans l'exemple du calcul de la part de population pauvre, les zones comportant des petits effectifs pourraient à tort apparaître distinctement dans la carte lissée. Peu peuplées, elles n'apparaîtraient pas sur une carte représentant les données brutes. Ainsi, en ne prenant pas garde à ce phénomène, on pourrait mécaniquement donner l'impression, faussée, que tous les territoires seraient peuplés.

8.2.4 Application utilisant un lissage quantile

Le lissage décrit jusqu'à présent est un lissage moyen, dans le sens où il est fondé sur des calculs locaux de moyennes. Dans l'article de BRUNSDON et al. 2002, les auteurs étendent cette notion, pour définir des statistiques locales fondées sur des quantiles (médiane, déciles ...). Ces indicateurs sont réputés, dans l'analyse exploratoire des données "classique", pour être moins sensibles aux valeurs extrêmes. Le lissage quantile permet surtout de calculer des indicateurs qui enrichissent considérablement l'analyse de certaines variables, notamment les variables de revenus.

Les quatre vignettes de la figure 8.9 représentent plusieurs indicateurs lissés calculés à partir du niveau de vie (source *Insee-DGFiP-Cnaf-Cnav-CCMSA, Fichier localisé social et fiscal 2012*), c'est-à-dire le revenu disponible d'un ménage divisé par le nombre d'unités de consommation de ce ménage.

Les cartes de la figure 8.9 sont centrées sur Paris, et incluent la "petite couronne". Le lissage est réalisé sur des carreaux de 50m, avec une bande passante de 400m. N'ont été retenus pour la visualisation que les carreaux pour lesquels le nombre d'observations (de ménages) ayant contribué à l'estimation est strictement supérieur à 50.

La carte 8.9a représente le niveau de vie médian. On retrouve, à l'Ouest, des zones où les habitants sont beaucoup plus aisés. Les cartes 8.9b et 8.9c représentent le 1er décile et le 9e décile du niveau de vie. La carte 8.9d représente le rapport interdécile, ratio entre le 9e et le 1er décile, et apporte un éclairage complémentaire. Exprimé sans unité, il s'agit du niveau de vie minimal des 10 % les plus riches rapporté au niveau de vie maximal des 10 % les plus pauvres. Il met en évidence l'écart entre le haut et le bas de la distribution : c'est une des mesures de l'inégalité de cette distribution. Dans les zones situées à l'Ouest, les rapports interdéciles sont très élevés : dans ces quartiers, cohabitent des populations dont le niveau de vie est très élevé, avec des populations dont le niveau de vie est beaucoup plus faible.

8.3 Mise en œuvre avec R

Avec R, plusieurs packages permettent de réaliser des lissages. Nous détaillerons ci-dessous la mise en œuvre pratique en utilisant les packages *spatstat* et *btb*, appliqués à des données concernant l'Île de la Réunion. Les données utilisées en exemple sont le `dataframe` *reunion.Rdata* fourni dans le package *btb*. Un aperçu de ce `dataframe` est donné sur la figure 8.10.

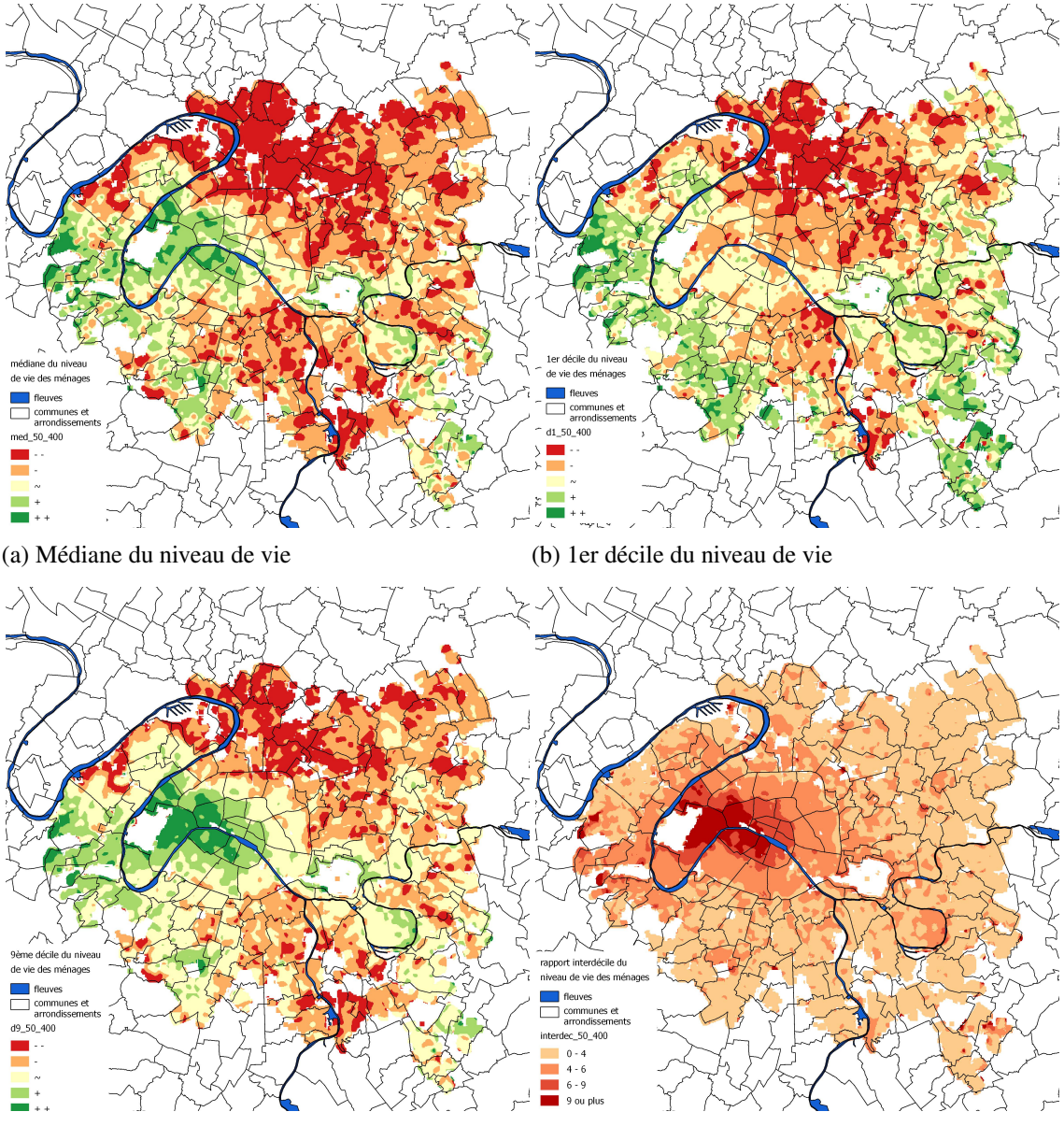
Il s'agit de données carroyées à 200 mètres, téléchargeables sur `insee.fr`. La source est *Insee, Revenus Fiscaux Localises (RFL) au 31 décembre 2010 et Taxe d'habitation (TH) au 1er janvier 2011*.

Les variables sont ainsi définies :

- x : longitude (système de projection : WGS 84 / UTM zone 40S, code EPSG : 32740)
- y : latitude (système de projection : WGS 84 / UTM zone 40S, code EPSG : 32740)
- houhold : nombre de ménages
- phouhold : nombre de ménages pauvres (définition pauvreté à 60 %)

8.3.1 Sous R, avec le package spatstat

Le package R appelé *spatstat* est un package très complet dédié à l'analyse des processus de points spatiaux. Il est disponible sur le site du CRAN à l'adresse suivante : `https://CRAN`.



(a) Médiane du niveau de vie

(b) 1er décile du niveau de vie

(c) 9ème décile du niveau de vie

(d) Rapport interdécile du niveau de vie

FIGURE 8.9 – Distribution du niveau de vie
Source : *Insee-DGFiP-Cnaf-Cnav-CCMSA, Fichier localisé social et fiscal 2012*
Note : Les carreaux représentés contiennent plus de 11 ménages. Pour les cartes représentant des niveaux de revenus, les marqueurs "++", "+", "∼", "-" et "- -" correspondent à des valeurs respectivement très élevées, élevées, moyennes, basses ou très basses pour l'indicateur considéré. Ils ont été utilisés pour des questions de non-profilage de la population

	x	y	houhold	phouhold
1	359500	7634300	5.0693069	2.37623762
2	359500	7634500	26.9306931	12.62376238
3	355900	7634500	15.0000000	4.00000000
4	356100	7634500	39.0000000	20.00000000
5	356300	7634500	41.6428571	15.14285714
6	356500	7634500	2.3571429	0.85714286
7	359700	7634500	11.4210526	0.00000000
8	359700	7634700	2.5789474	0.00000000
9	359900	7634500	12.0000000	6.00000000
10	355700	7634700	1.0243902	0.00000000
11	355700	7635100	1.3658537	0.00000000
12	355700	7635300	11.6097561	0.00000000
13	355900	7634700	20.0000000	7.00000000
14	356100	7634700	131.0000000	71.00000000
15	356300	7634700	110.0000000	58.00000000

FIGURE 8.10 – Les 15 premières lignes du `data.frame` *reunion.Rdata* du package *btb*
Source : *Insee, Revenus Fiscaux Localises (RFL) au 31 décembre 2010 et Taxe d'habitation (TH) au 1er janvier 2011*
Champ : Île de la Réunion

`R-project.org/package=spatstat`

La fonction `density.ppp` disponible dans le package *spatstat* permet d'effectuer le lissage des données. L'utilisation de cette fonction nécessite en entrée l'utilisation d'un objet au format *.ppp*. Afin d'utiliser cette fonction, les coordonnées x et y du `data.frame` de départ doivent être converties au format *.ppp*.

```
#lissage de la variable houhold (nombre de ménages) avec Spatstat

library(spatstat)
library(btb) #uniquement pour le dataframe reunion
data(reunion)

# agrégation et suppression des doublons sur les coordonnées
base_temp <- aggregate(houhold ~ x+y, reunion, sum)

# transformation des x,y en objet .ppp
base.ppp = spatstat::ppp(base_temp$x, base_temp$y,
               c(min(base_temp$x), max(base_temp$x)),
               c(min(base_temp$y), max(base_temp$y)) )

#appel de la fonction density.ppp
#le parametre sigma correspond à h/2 avec h la bande passante
densite <- spatstat::density.ppp (base.ppp, sigma = 200, weights=base_temp$
    houhold )

#affichage de la carte
plot(densite, main = "Lissage spatstat, rayon défini par utilisateur")
```

8.3.2 Sous R, avec le package btb

Le package *btb* ("beyond the border") [1], en ligne sur le site du CRAN à l'adresse suivante : `https://CRAN.R-project.org/package=btb`, propose des fonctions dédiées à l'analyse urbaine. Il met en œuvre une estimation de densité par la méthode KDE (*kernel density estimator*), c'est-à-dire une méthode par noyau. Le noyau utilisé est un noyau quadratique.

Dans l'estimation réalisée par le package, l'effet des frontières est pris en compte dans la fonction de lissage `kernelSmoothing` *via* la correction de Diggle (DIGGLE 2013). Cette correction permet notamment de traiter le cas de frontières qui représentent des limites géographiques (des côtes maritimes par exemple). À l'intérieur de la zone d'observation, l'intensité est non nulle. À l'extérieur de la zone d'observation, elle est nulle. La méthode implémentée est conservative (grâce à une normalisation) : avant et après lissage, le nombre de points observés est identique.

Les temps de calcul ont été fortement réduits, et ce de plusieurs manières :
- en codant en C++ les méthodes les plus chronophages ;
- en se restreignant, pour chaque point, à une fenêtre d'observation autour de ce point, permettant de limiter le nombre d'opérations (calculs de distances) à effectuer.

1. Il y aura prochainement une version du package *btb* qui sera adaptée au nouveau package *sf* (simple features)

```
#lissage avec btb : calcul de la part de ménages pauvres
#on lisse séparément le numérateur (nombre de ménages pauvres), et le dé
    nominateur (nombre total de ménages)

library(btb)

#chargement des données
data(reunion)

#lissage
#définition des paramètres
pas <- 200 #carreau de 200m de cote
rayon <- 400 #bande passante de 400m

#appel de la fonction de lissage
#la fonction lisse automatiquement l'ensemble des variables contenues dans
    la base
#ici on lisse phouhold et  houhold
dfLisse <- btb::kernelSmoothing(dfObservations = reunion, iCellSize = pas,
                                          iBandwidth = rayon, sEPSG="32740")

#taux de ménages pauvres : ratio des variables lissées
dfLisse$txmenpa = 100 * dfLisse$phouhold / dfLisse$houhold

#aperçu dans R
library(sp)
library(cartography)
#affichage de la carte
cartography::choroLayer(dfLisse, var = "txmenpa", nclass = 5, method = "
    fisher-jenks", border = NA, legend.pos = "topright", legend.title.txt =
     "txmenpa (%)")

#ajout du titre et d'un contour
cartography::layoutLayer(title = "La Réunion : taux de ménages pauvres",
                         sources = "",
                         author = "",
                         scale = NULL,
                         frame = TRUE,
                         col = "black",
                         coltitle = "white")
```

La carte obtenue est représenté sur la figure 8.11.

L'utilisateur peut également exporter le résultat au format shapefile pour le retravailler ensuite dans un SIG.

```
#export au format shapefile

rgdal::writeOGR(as(dfLisse, 'Spatial'), "txmenpauvre.shp", "txmenpauvre",
    driver = "ESRI Shapefile")
```

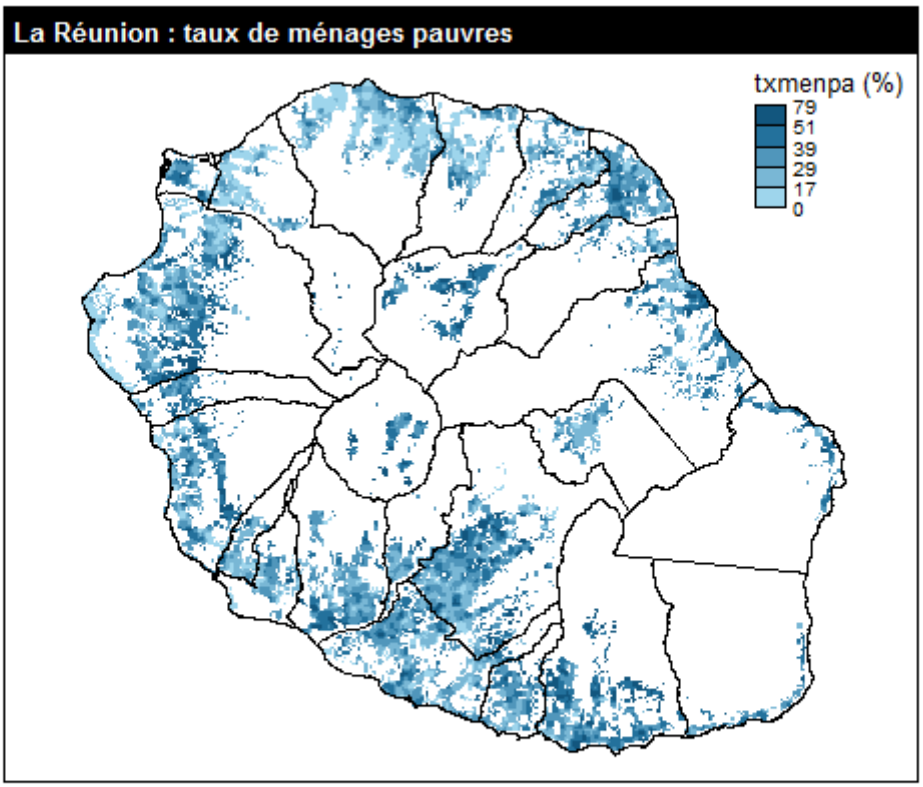


FIGURE 8.11 – Taux de ménages pauvres à la Réunion après lissage
Source : *Insee, Revenus Fiscaux Localises au 31/12/2010 et Taxe d'habitation au 01/01/2011*
Note : ratio du nombre de ménages pauvres lissé et du nombre total de ménages lissé. Les contours noirs représentent les limites communales

Le package *btb* permet également d'utiliser le lissage quantile, décrit plus haut. Il suffit de spécifier comme paramètre `vQuantiles` le vecteur des quantiles à calculer. Par exemple *c(0.1, 0.25, 0.5)* retournera le premier décile, le premier quartile et la médiane de chacune des variables du `data.frame` en entrée.

```
# lissage quantile
library(btb)
data(reunion)

# définition des paramètres
pas <- 200
rayon <- 400

# appel de la fonction de lissage
dfLisse_quantile<- btb::kernelSmoothing(dfObservations = reunion,
                                        iCellSize = pas,
                                        iBandwidth = rayon,
                                        vQuantiles = c(0.1, 0.5, 0.9),
                                        sEPSG="32740")

#export vers QGIS
rgdal::writeOGR(as(dfLisse_quantile, 'Spatial'), "lissage_quantile.shp", "
    lissage_quantile",
                                driver = "ESRI Shapefile")
```

R Le package *btb* propose par défaut une grille automatique de carreaux. Il est également possible d'appeler la fonction de lissage en utilisant une grille au choix de l'utilisateur. Dans ce cas, l'utilisateur doit disposer d'un `data.frame` composé de deux colonnes x et y, qui correspondent aux coordonnées des centroïdes souhaités.

```
# fonction de lissage avec grille au choix de l'utilisateur
kernelSmoothing(dfObservations, iCellSize, iBandwidth, dfCentroids)
```

8.3.3 Tests de bande passante optimale

Dans R, plusieurs méthodes proposant de calculer une bande passante "optimale" sont implémentées, selon différents critères. L'objectif est généralement de minimiser une mesure d'erreur. Dans *spatstat* par exemple, il existe les quatre fonctions suivantes : `bw.diggle`, `bw.ppl`, `bw.frac` et `bw.scott`.

Avec la fonction `bw.diggle` de *spatstat*

La fonction `bw.diggle` de *spatstat* choisit une bande passante qui minimise un critère M(σ) basé sur l'erreur quadratique moyenne (en anglais MSE pour *Mean Square Error*) de l'estimateur.

Le graphique de la figure 8.12 représente le critère M(σ) que l'on souhaite minimiser. Pour obtenir la valeur σ, il s'agit de repérer sur l'axe des abscisses la valeur qui correspond à la valeur minimale en ordonnées.

Pour plus de détails, voir `https://www.rdocumentation.org/packages/spatstat/versions/1.49-0/topics/bw.diggle`.

```
#on réutilise base.ppp créée plus haut
# test bw.diggle de bande passante optimale
bw_diggle <- spatstat::bw.diggle(base.ppp)
plot(bw_diggle, main = "cross validation")

#appel de density.ppp avec la bande passante calculée automatiquement
densite_optim <- spatstat::density.ppp(base.ppp,bw_diggle, weights=base_
    temp$houhold)
```

On obtient :

```
bw_diggle
##    sigma
## 141.9445
```

Avec les paramètres par défaut, la valeur proposée pour σ est de 142 mètres soit 284 mètres pour la bande passante h (σ = h/2 , voir la documentation du package).

Avec la fonction `bw.ppl` de *spatstat*

La bande passante est choisie en calculant un estimateur de maximum de vraisemblance, en utilisant une méthode de validation croisée (*likelihood cross-validation criterion*). On itère les calculs : à chaque fois, on ne travaille que sur $n-1$ observations puis on valide le modèle sur l'observation qui avait été écartée. On répète cela n fois.

Le graphique ci-dessous représente le critère CV(σ) que l'on souhaite minimiser. Pour obtenir la valeur σ, il s'agit de repérer sur l'axe des abscisses la valeur qui correspond à la valeur maximale en ordonnées.

Pour plus de détails, voir `https://rdrr.io/cran/spatstat/man/bw.ppl.html`.

```
#on réutilise base.ppp créée plus haut

# test bw.ppl de bande passante optimale
bw_ppl <- spatstat::bw.ppl(base.ppp)
plot(bw_ppl, main = "bw.ppl")
```

On obtient :

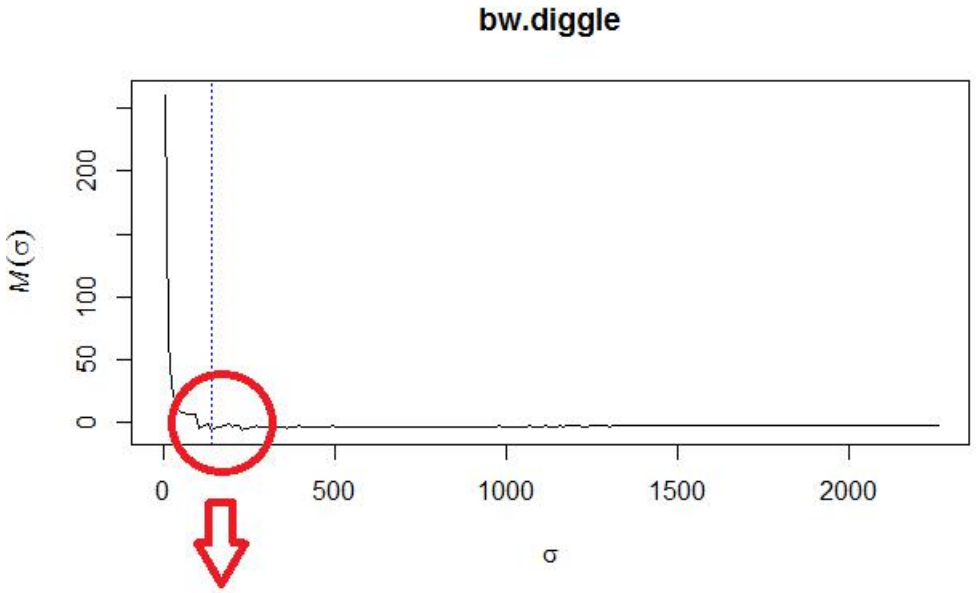
```
bw_ppl
##    sigma
## 286.0097
```

Avec les paramètres par défaut, la valeur proposée pour la valeur de σ est de 286 mètres.

Avec la fonction `bw.frac` de *spatstat*

Cette méthode sélectionne une bande passante qui repose uniquement sur la géométrie de la fenêtre d'observation.

La bande passante est un quantile (que l'on spécifie) de la distance entre deux points indépendants, pris au hasard dans la fenêtre. Par défaut, c'est le premier quartile de la distribution qui est utilisé. Si on note CDF(r) la fonction de distribution cumulée de la distance entre deux points indépendants pris au hasard et uniformément distribués dans la fenêtre, alors la valeur qui



En zoomant autour de la valeur proposée :

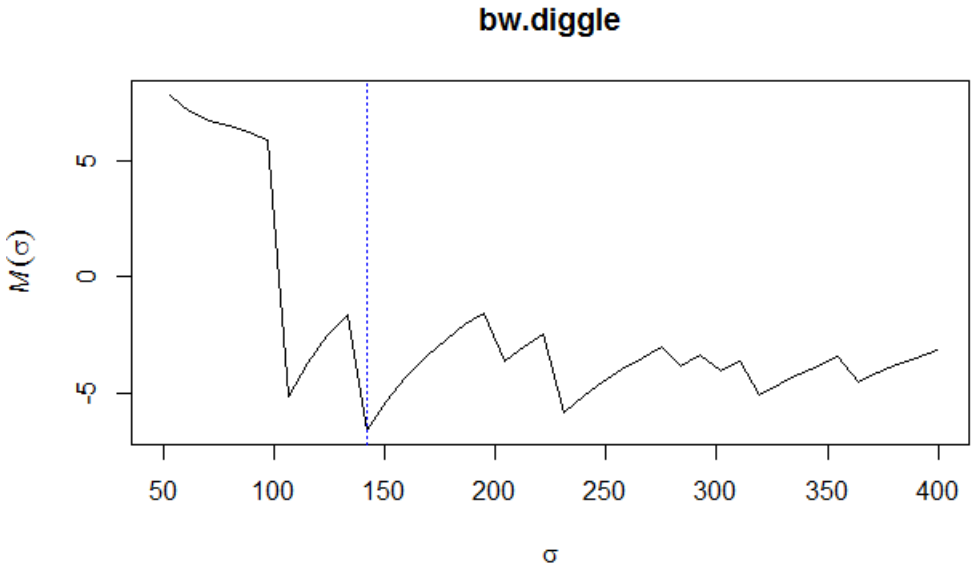
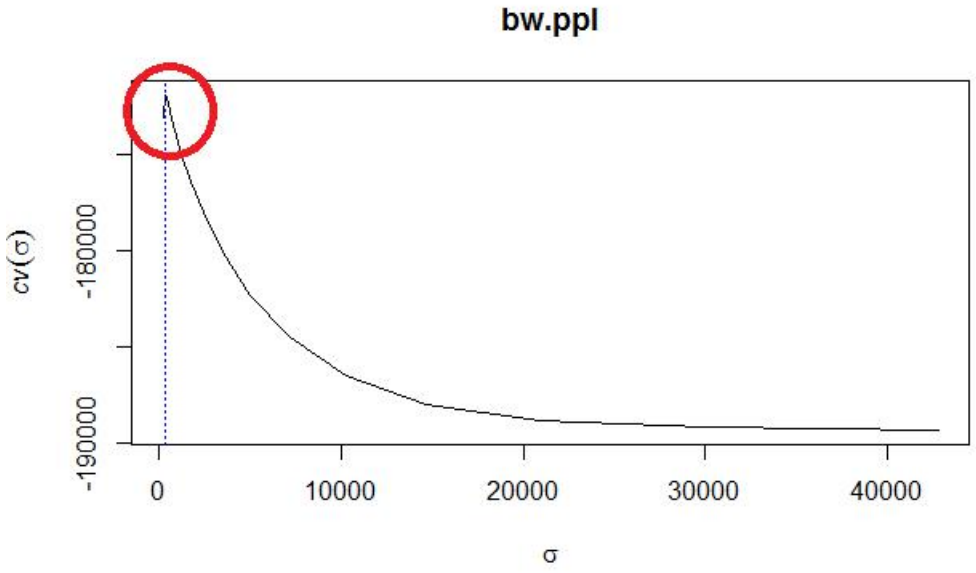


FIGURE 8.12 – Le critère M(σ) obtenu par la fonction `bw.diggle` du package *spatstat* de R
Source : *Insee, Revenus Fiscaux Localises (RFL) au 31 décembre 2010 et Taxe d'habitation (TH) au 1er janvier 2011*
Champ : Île de la Réunion



En zoomant autour de la valeur proposée :

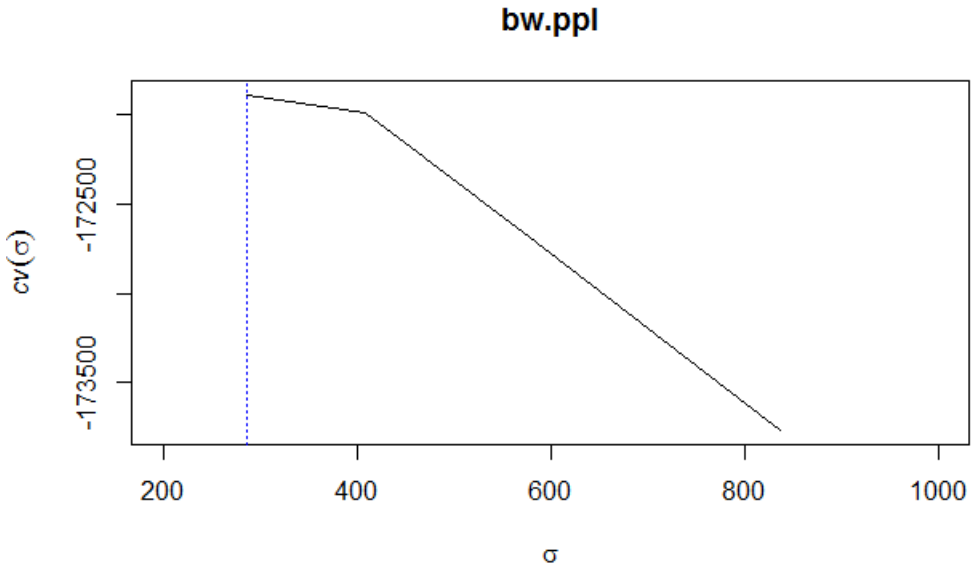


FIGURE 8.13 – Le critère CV(σ) obtenu par la fonction `bw.ppl` du package *spatstat* de R
Source : *Insee, Revenus Fiscaux Localises (RFL) au 31 décembre 2010 et Taxe d'habitation (TH) au 1er janvier 2011*
Champ : Île de la Réunion

est retournée est le quantile avec la probabilité f. Alors la bande passante est la valeur r telle que CDF(r)=f. En premier, l'algorithme calcule la fonction de distribution cumulée CDF(r) avec la fonction `distcdf` du package *spatstat*. Cette fonction permet de calculer la fonction CDF(r) = P(T<=r) de la distance euclidienne T=|X_1-X_2| entre deux points indépendants pris au hasard X_1 et X_2. Ensuite, on cherche le plus petit nombre r tel que CDF(r) $\geq$ f.

Le graphique ci-dessous représente la fonction CDF(r). Pour obtenir la bande passante, on lit la valeur des abscisses r telle que CDF(r) = 0.25 (par défaut c'est le premier quartile qui est utilisé).

Pour plus de détails voir `https://www.rdocumentation.org/packages/spatstat/versions/1.48-0/topics/bw.frac`.

```
#on réutilise base.ppp créée plus haut

# test bw.frac de bande passante optimale
bw_frac<-  spatstat::bw.frac(base.ppp)
plot(bw_frac, main = "bw.frac")
```

On obtient :

```
bw_frac
## [1] 19747.02
```

Avec les paramètres par défaut, la valeur proposée pour la valeur de σ est de 19747 mètres.

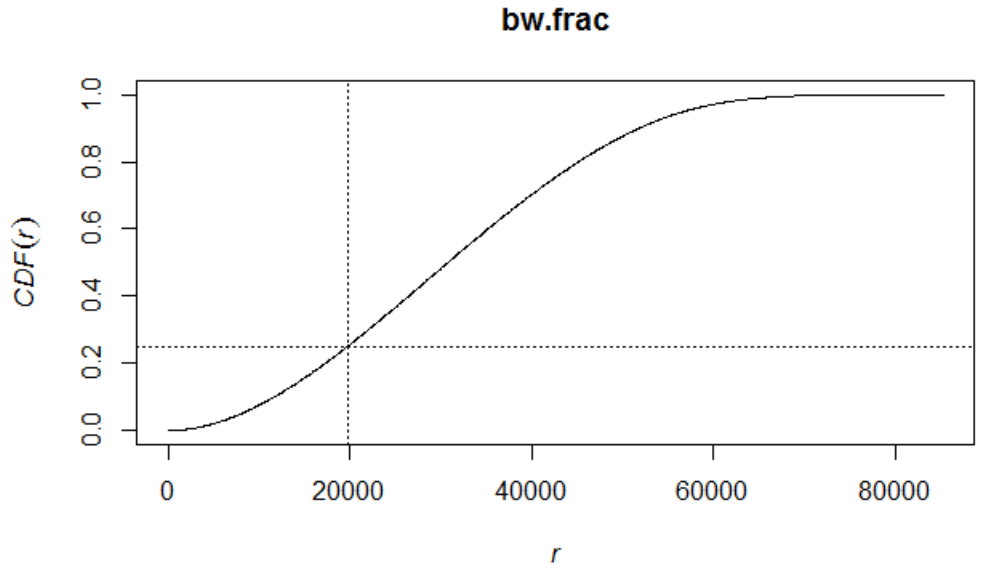


FIGURE 8.14 – La fonction de distribution cumulée CDF(r) obtenue par la fonction `bw.frac` du package *spatstat* de R
Source : *Insee, Revenus Fiscaux Localises (RFL) au 31 décembre 2010 et Taxe d'habitation (TH) au 1er janvier 2011*
Champ : Île de la Réunion

Avec la fonction `bw.scott` de *spatstat*

Cette fonction est basée sur la «règle de Scott» (voir SCOTT 1992). Il s'agit de supposer que l'échantillon est distribué selon une loi normale. Dans ce cas, on obtient un estimateur de la bande passante, en minimisant une erreur appelée «erreur moyenne quadratique intégrée». Dans la formule de l'estimateur intervient notamment l'écart-type de l'échantillon.

Le résultat obtenu est un vecteur composé de deux valeurs : les bandes passantes proposées dans la direction des x et des y.

```
#on reutilise base.ppp créée plus haut

#test bw.scott de bande passante optimale
bw_scott <- spatstat::bw.scott(base.ppp)
```

On obtient :

```
bw_scott
## [1] 2973.548 3455.256
```

Avec les paramètres par défaut, la valeur proposée est le couple (2974 ; 3455) : 2974 mètres dans la direction des x et 3455 dans la direction des y. D'après la documentation du package, la valeur proposée par ce test est généralement plus élevée que celle fournie par `bw.diggle`.

Synthèse des résultats obtenus

fonction	σ(en mètres)
bw.diggle	142
bw.ppl	286
bw.frac	19747
bw.scott	2974 (x) et 3455 (y)

TABLE 8.1 – Bandes passantes "optimales" ($h = 2\sigma$) obtenues par les fonctions du package *spatstat*
Source : *Insee, Revenus Fiscaux Localises (RFL) au 31 décembre 2010 et Taxe d'habitation (TH) au 1er janvier 2011*
Champ : Île de la Réunion

Dans cet exemple, les résultats diffèrent parfois beaucoup selon les méthodes. Les écarts sont exacerbés par la répartition atypique de la population sur l'île de la Réunion, quasi-exclusivement située sur le littoral.

Conclusion

Derrière la qualité esthétique des cartes lissées se cache néanmoins un piège majeur. Par construction, les méthodes de lissage atténuent les ruptures et les frontières et induisent des représentations continues des phénomènes géographiques. Les cartes lissées font donc apparaître localement de l'autocorrélation spatiale. Deux points proches par rapport au rayon de lissage ont mécaniquement des caractéristiques comparables dans ce type d'analyse. De ce fait, commenter à partir d'une carte lissée des phénomènes géographiques dont l'ampleur spatiale est de l'ordre du rayon de lissage n'a guère de sens. Intuitivement, cela revient à commenter l'homogénéité observée au sein des unités spatiales d'une carte choroplèthe. Autrement dit, le rayon de lissage (la bande

passante) définit implicitement une maille minimale de restitution de l'information. En corollaire de ces remarques, il est primordial de commenter uniquement des phénomènes dont l'ordre de grandeur est très supérieur au rayon de lissage.

Références - Chapitre 8

BADDELEY, A. et al. (2015a). *Spatial Point Patterns : Methodology and Applications with R*. CRC Press.

BRUNSDON, C. et al. (2002). « Geographically weighted summary statistics : a framework for localised exploratory data analysis ». *Computers, Environment and Urban Systems*.

DIGGLE, Peter J. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. CRC Press.

FLOCH, J.M. (2012). « Détection des disparités socio-économiques - L'apport de la statistique spatiale ». *Documents de Travail Insee*.

PALSKY, Gilles (1991). « La cartographie statistique de la population au XIXe siecle ». *Espace, populations, sociétés* 9.3, p. 451–458.

SCOTT, D.W. (1992). *Multivariate Density Estimation : Theory, Practice, and Visualization.* New York, Chichester : Wiley.

9. Régression géographiquement pondérée

MARIE-PIERRE DE BELLEFON, JEAN-MICHEL FLOCH
Insee

Résumé

La régression géographiquement pondérée (RGP) répond au constat qu'un modèle de régression estimé sur l'ensemble d'un territoire d'intérêt peut ne pas appréhender de façon adéquate les variations locales. Son principe, assez simple, consiste en l'estimation de modèles locaux par les moindres carrés, chaque observation étant pondérée par une fonction décroissante de sa distance au point d'estimation. La réunion de ces modèles locaux permet la construction d'un modèle global aux propriétés spécifiques. La RGP permet, notamment à l'aide de représentations cartographiques associées, de repérer où les coefficients locaux s'écartent le plus des coefficients globaux, de construire des tests permettant d'apprécier si le phénomène est non stationnaire et de caractériser la non stationnarité. La méthode est présentée à partir de l'exemple d'un modèle des prix hédoniques (prix des logements anciens à Lyon). Nous montrons comment déterminer de façon optimale le rayon du disque sur lequel seront effectuées les régressions locales et présentons les résultats d'estimation, les méthodes d'estimation robustes et les tests de non-stationnarité des coefficients. En complément de cette utilisation descriptive, nous présentons une approche plus prédictive, montrant comment la prise en compte de la non-stationnarité permet d'améliorer un estimateur sur un domaine spatial. L'exemple est construit à partir d'un modèle liant la population pauvre et le nombre de bénéficiaires de la couverture maladie universelle complémentaire (CMU-C) à Rennes.

R La lecture préalable du chapitre 3 : "Indices d'autocorrélation spatiale" est recommandée.

9.1 Pourquoi utiliser une régression géographiquement pondérée ?

Pour identifier la nature des relations entre les variables, la régression linéaire modélise la variable dépendante y comme une fonction linéaire des variables explicatives $x_1, ..., x_p$. Si l'on dispose de n observations, le modèle s'écrit :

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \varepsilon_i \tag{9.1}$$

avec β_0, β_1, ..., β_p les paramètres et ε_1, ε_2,..., ε_n les termes d'erreur. Dans ce modèle, les coefficients β_k sont considérés comme identiques sur toute la zone d'étude. Cependant, cette hypothèse d'uniformité spatiale de l'effet des variables explicatives sur la variable dépendante est souvent irréaliste (BRUNSDON et al. 1996). Si les paramètres varient significativement dans l'espace, un estimateur global occultera la richesse géographique du phénomène étudié.

L'hétérogénéité spatiale correspond à cette variabilité spatiale des paramètres ou de la forme fonctionnelle du modèle. Lorsque l'on dispose d'une bonne connaissance du territoire d'intérêt, elle est souvent traitée dans la littérature empirique en ajoutant des indicatrices de zones géographiques dans le modèle (éventuellement croisées avec chaque variable explicative), et en estimant le modèle pour différentes zones ou en conduisant des tests de stabilité géographique des paramètres (dits de Chow). Lorsque le nombre de ces zones géographiques augmente, ce traitement diminue néanmoins le nombre de degrés de liberté et donc la précision des estimateurs.

On peut également utiliser des régressions locales dont l'application spatiale est la régression géographique pondérée (GWR, Geographically Weighted Regression, BRUNSDON et al. 1996). À travers l'exemple de l'étude des prix de l'immobilier à Lyon, nous présentons l'intérêt d'effectuer une régression géographiquement pondérée (exemple 9.1) et la façon dont elle peut être mise en œuvre (exemple 9.2).

Des méthodes plus complexes venues du monde de la géographie ont été développées (LE GALLO 2004), mais elles restent en grande partie descriptives et exploratoires (notamment à travers des représentations graphiques), car leur comportement théorique n'est pas complètement connu, notamment en ce qui concerne la convergence et la prise en compte des ruptures géographiques.

■ **Exemple 9.1 — Utilisation d'un modèle hédonique pour étudier les prix de l'immobilier lyonnais.** Cartographier les variations des prix de l'immobilier permet de déduire de façon générale que les prix ont tendance à être plus élevés dans le centre qu'en périphérie (figure 9.1). Cependant, ces prix élevés s'expliquent peut-être par une meilleure qualité des logements vendus dans le centre. Le modèle hédonique a pour objectif d'**isoler l'effet de la localisation sur les prix.** Le principe de cette méthode est que le prix d'un bien est une combinaison des prix de ses différents attributs.

$$y_i = \beta_0 + \sum_{k}^{p} \beta_k x_{ik} + \varepsilon_i \tag{9.2}$$

avec x_{ik} la caractéristique k du bien i, β_k le coefficient associé à cette caractéristique et p le nombre de variables explicatives.

Les hypothèses sous-jacentes au modèle hédonique sont que les vendeurs et les acheteurs sont des agents individuels, sans pouvoir de marché, et qu'il s'agit d'une situation de concurrence parfaite. Le coefficient de la régression hédonique associé à une caractéristique informe sur la valeur que les acheteurs **à l'équilibre à un instant donné** accorderaient à **une augmentation de la quantité de cette caractéristique.**

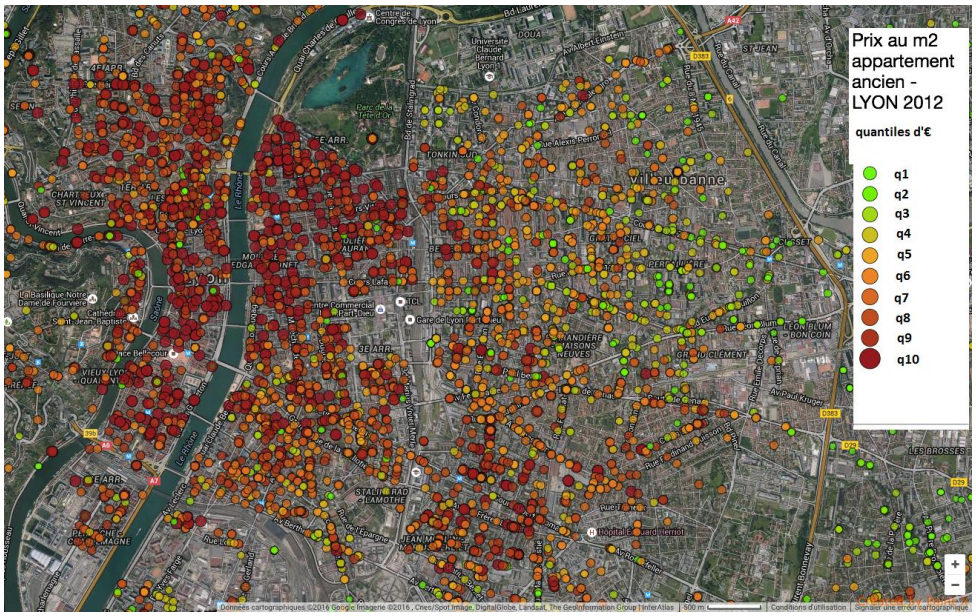


FIGURE 9.1 – Prix de vente au m² d'un appartement ancien - 2012
Source : *base PERVAL*
Champ : agglomération lyonnaise

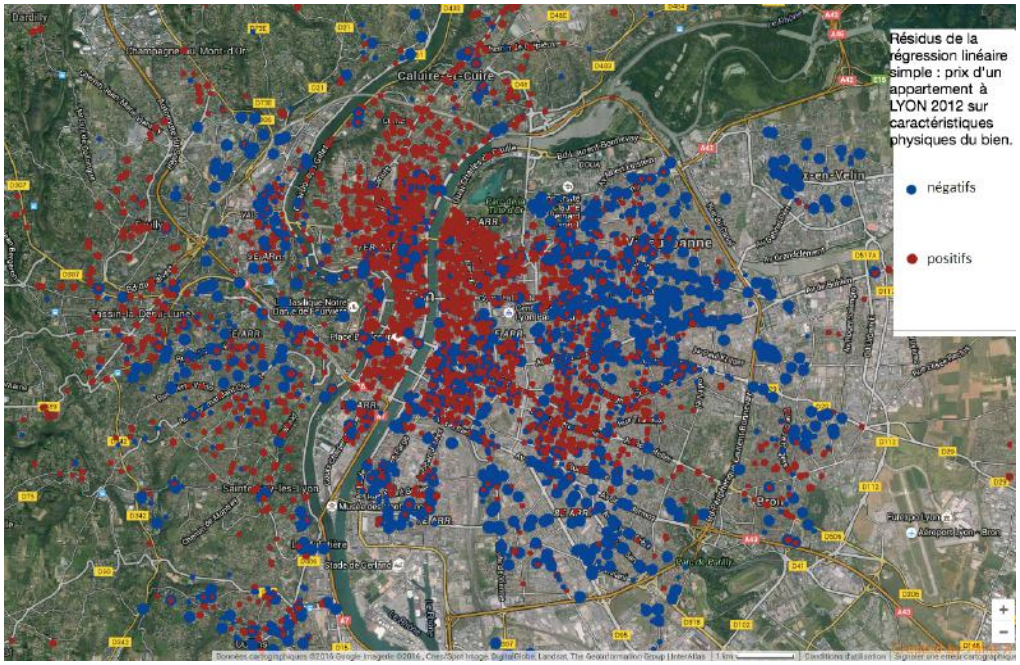


FIGURE 9.2 – Résidus de la régression hédonique du prix sur les caractéristiques du bien
Source : *base PERVAL*
Champ : agglomération lyonnaise

La figure 9.2 représente les résidus de la régression hédonique des prix des appartements sur leurs caractéristiques physiques. Ces résidus ne sont pas distribués aléatoirement dans l'espace (l'hypothèse nulle du test de Moran est rejetée). Le I de Moran de la distribution des résidus est positif, signe d'une corrélation spatiale positive des résidus. **L'hypothèse de stationnarité spatiale de la relation entre prix et caractéristique du bien n'est donc pas valide**, ce qui indique l'existence d'un phénomène d'**hétérogénéité spatiale**.
Comme cela a été vu précédemment, pour tenir compte de la variation des paramètres du modèle avec la localisation, une méthode couramment utilisée consiste à introduire des indicatrices de localisation comme paramètres explicatifs. Considérons à titre d'exemple les prix au m² des appartements à Lyon en 2012 et étudions l'impact sur ces prix de la période de construction de ces appartements. Pour cela nous introduisons la variable indicatrice qui indique pour chaque appartement si celui-ci a été construit entre 1992 et 2000 (indicatrice = 1) plutôt qu'entre 1948 et 1969 (indicatrice = 0). La table 9.1 indique la valeur des coefficients associés à cette indicatrice pour chaque arrondissement lyonnais.

Arrondissement	Coefficient de la régression	Significativité
1er	1,1511	.
2ème	1,1499	.
3ème	**1,1481**	***
4ème	1,1860	**
5ème	1,4909	***
6ème	**1,3085**	***
7ème	1,1897	***
8ème	1,1487	***
9ème	1,1981	***

TABLE 9.1 – Significativité des coefficients de régression associés à une époque de construction récente
*, **, *** indiquent la significativité aux seuils de 10, 5 et 1% **Source :** *base PERVAL*
Champ : agglomération lyonnaise

La valeur et la significativité des coefficients varient en fonction des arrondissements (table 9.1). Les acheteurs valorisent donc différemment l'époque de construction d'un logement suivant sa localisation. Cependant, pourquoi les frontières qui définissent les changements de modèle correspondraient-elles nécessairement à des contours administratifs ? **La régression géographiquement pondérée permet de répondre à cette question et d'étudier un phénomène qui varie continûment dans l'espace.**

■

9.2 La régression géographiquement pondérée

9.2.1 Un modèle à coefficients variables

La régression géographiquement pondérée appartient à la catégorie des modèles à coefficients variables. Les coefficients de la régression ne sont pas fixes : ils dépendent des coordonnées géographiques des observations. Dit autrement : **les coefficients des paramètres explicatifs forment des surfaces continues qu'on estime en certains points de l'espace**

$$y_i = \beta_0(u_i, v_i) + \sum_k^p \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \tag{9.3}$$

avec (u_i,v_i) les coordonnées géographiques.

9.2.2 Comment estimer le modèle ?

Pour estimer le modèle, on utilise l'hypothèse suivante : **plus deux observations sont proches dans l'espace, plus l'influence des variables explicatives sur la variable dépendante est proche ; c'est à dire, plus les coefficients des paramètres explicatifs de la régression sont proches.** Par conséquent, pour estimer le modèle à coefficients variables au point i, on souhaite utiliser le modèle à coefficients fixes en incluant dans la régression uniquement les observations proches de i. Or plus on inclut de points dans l'échantillon, plus la variance est faible, mais plus le biais est élevé. La solution consiste donc à **diminuer l'importance des observations les plus éloignées en accordant à chaque observation un poids décroissant avec la distance au point d'intérêt.**

Le modèle à estimer est le suivant :

$$\mathbf{Y} = (\beta \otimes \mathbf{X})\mathbf{1} + \varepsilon \tag{9.4}$$

$\mathbf{Y}$: vecteur $n \times 1$ de la variable dépendante.
$\mathbf{X}$: matrice $n \times (p+1)$ des p variables explicatives + la constante.
$\mathbf{1}$: vecteur $(p+1) \times 1$ de 1

Les coefficients β du modèle peuvent être exprimés sous forme matricielle :

$$\beta = \begin{bmatrix} \beta_0(u_1,v_1) & \dots & \beta_p(u_1,v_1) \\ \beta_0(u_j,v_j) & \dots & \beta_p(u_j,v_j) \\ \beta_0(u_n,v_n) & \dots & \beta_p(u_n,v_n) \end{bmatrix} \tag{9.5}$$

L'opérateur $\otimes$ multiplie chaque élément de la matrice des coefficients β par l'élément correspondant de la matrice X des caractéristiques des observations.

Pour accorder un poids décroissant aux observations en fonction de leur distance au point d'intérêt, on effectue une estimation par moindres carrés pondérés, la pondération étant régie par la matrice de poids $W_{(u_i,v_i)}$. Les paramètres régissant la construction de cette matrice sont détaillés dans la section 9.2.3.

Conformément au principe des moindres carrés pondérés, les coefficients $\hat{\beta}(u_i,v_i)$ au point de coordonnées géographiques (u_i,v_i) minimisent la somme 9.6 :

$$\sum_{j=1}^{n} w_j(i)(y_j - \beta_0(u_i,v_i) - \beta_1(u_i,v_i)x_{j1} - \dots - \beta_p(u_i,v_i)x_{jp})^2 \tag{9.6}$$

$$\hat{\beta}(u_i,v_i) = (\mathbf{X^T W_{(u_i,v_i)} X})^{-1}\mathbf{X^T W_{(u_i,v_i)} Y} \tag{9.7}$$

On peut écrire $\hat{\mathbf{Y}} = \mathbf{SY}$ avec S dénommée la "matrice chapeau" et définie par l'équation 9.8 : En notant $\mathbf{x_i^T} = (1 \quad x_{i1} \quad x_{i2} \quad \dots \quad x_{ip})$ la $i$ème colonne de $\mathbf{X}$, la matrice des variables explicatives, on a alors

$$\mathbf{S} = \begin{bmatrix} (\mathbf{x_1^T X^T W_{(u_1,v_1)} X})^{-1}\mathbf{X^T W_{(u_1,v_1)}} \\ \vdots \\ \mathbf{x_n^T X^T W_{(u_n,v_n)} X})^{-1}\mathbf{X^T W_{(u_n,v_n)}} \end{bmatrix} \tag{9.8}$$

Rappel : estimation par Moindres Carrés Ordinaires

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \tag{9.9}$$

$$\hat{\beta} = (\mathbf{X^T X})^{-1}\mathbf{X^T Y} \tag{9.10}$$

Y : vecteur nx1 de la variable dépendante.
X : matrice nx(p+1) des p variables explicatives + la constante.

9.2.3 Choisir les paramètres d'estimation

La matrice $W_{(u_i,v_i)}$ contient le poids de chaque observation en fonction de sa distance au point i de coordonnées (u_i, v_i) (figure 9.3). On suppose que les observations proches du point i exercent plus d'influence sur les paramètres estimés au lieu i que les observations plus lointaines. Le poids des observations est donc décroissant avec la distance au point i. Il existe plusieurs manières de spécifier cette décroissance. Nous présentons ici les principaux paramètres de décroissance.

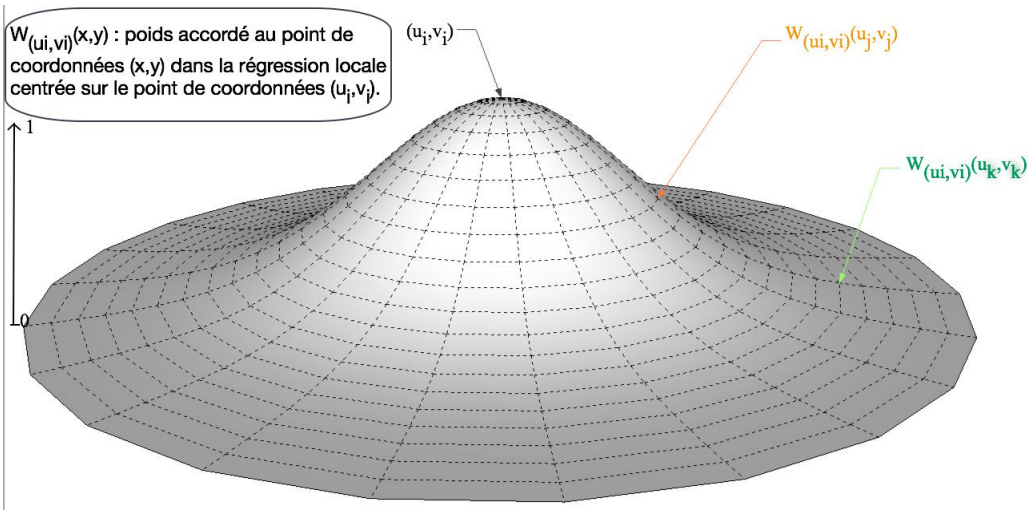


FIGURE 9.3 – Représentation graphique de la matrice W

La décroissance du poids de chaque observation avec la distance au point d'origine est déterminée par une **fonction de noyau**. Les paramètres clés de la fonction de noyau sont :

— la forme du noyau ;
— le noyau fixe ou adaptatif ;
— la taille de la bande passante.

La forme du noyau

On peut distinguer les noyaux continus qui accordent un poids à toutes les observations (figure 9.4 ; table 9.2) des noyaux à support compact (figure 9.5 ; table 9.3) pour lesquels le poids des observations est nul au delà d'une certaine distance. Cependant, **la forme du noyau ne modifie que légèrement les résultats** (BRUNSDON et al. 1998).

— Choisir un noyau uniforme revient à effectuer une régression par moindres carrés ordinaires en chaque point.

Noyau uniforme	$w(d_{ij}) = 1$
Noyau gaussien	$w(d_{ij}) = exp(-\frac{1}{2}(\frac{d_{ij}}{h})^2)$
Noyau exponentiel	$w(d_{ij}) = exp(-\frac{1}{2}(\frac{\lvert d_{ij}\rvert}{h}))$

TABLE 9.2 – Expression fonctionnelle de noyaux continus

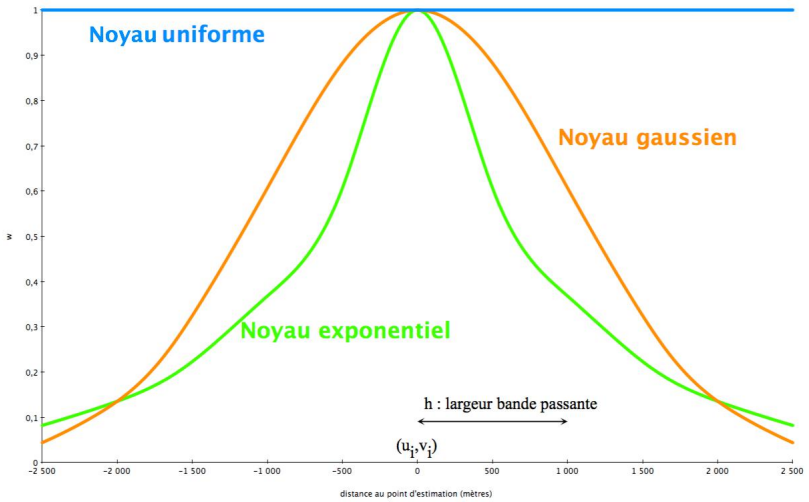


FIGURE 9.4 – Représentation graphique de noyaux continus

Noyau Box-Car	$w(d_{ij}) = 1$ si $\lvert d_{ij}\rvert < h$, 0 sinon
Noyau Bi-square	$w(d_{ij}) = (1-(\frac{d_{ij}}{h})^2)^2$ si $\lvert d_{ij}\rvert < h$, 0 sinon
Noyau Tri-cube	$w(d_{ij}) = (1-(\frac{\lvert d_{ij}\rvert}{h})^3)^3$ si $\lvert d_{ij}\rvert < h$, 0 sinon

TABLE 9.3 – Expression fonctionnelle de noyaux à support compact

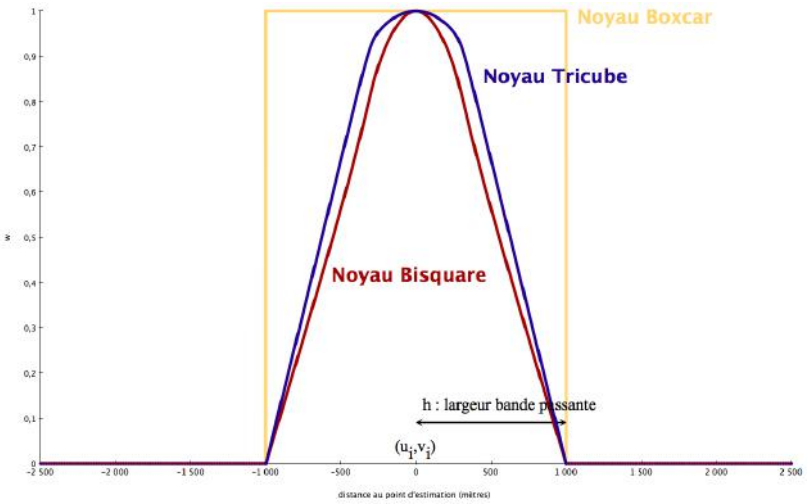


FIGURE 9.5 – Représentation graphique de noyaux à support compact

— Le noyau Box-Car traite un phénomène continu de façon discontinue.
— Les noyaux gaussiens et exponentiels pondèrent toutes les observations, avec un poids qui tend vers zéro avec la distance au point estimé.
— Les noyaux Bisquare et Tricube accordent également aux observations un poids décroissant avec la distance, mais ce poids est nul au delà d'une certaine distance h appelée *bande passante*.

⇒ **Le noyau Bisquare est à privilégier pour optimiser le temps de calcul.**

Noyau fixe ou adaptatif

Définition 9.2.1 — Noyau fixe. L'étendue du noyau est déterminée par la **distance** au point d'intérêt. Le noyau est identique en tout point de l'espace (figure 9.6).

Définition 9.2.2 — Noyau adaptatif. L'étendue du noyau est déterminée par le **nombre de voisins** du point d'intérêt. Plus la densité des observations est faible, moins le noyau est étendu (figure 9.7).

— Un noyau fixe est adapté à une répartition des données uniforme dans l'espace mais peu efficace dans le cas d'une répartition inhomogène. Son rayon doit être au moins égal à la distance entre le point le plus isolé et son premier voisin ce qui peut conduire à un nombre variable de points inclus dans la régression.
— Dans les zones peu denses, un noyau fixe trop petit inclura trop peu de points dans la régression. La variance sera plus élevée.
— Dans les zones très denses, un noyau fixe trop grand négligera les variations à une échelle fine. Le biais sera plus élevé.

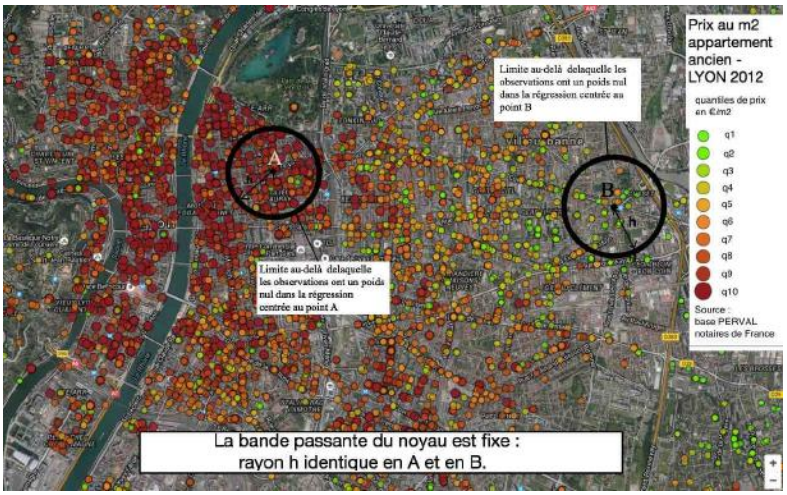


FIGURE 9.6 – Noyau fixe
Source : *base PERVAL*

Définition et choix de la bande passante

La bande passante est une distance au-delà de laquelle le poids des observations est considéré comme nul. **La valeur de la bande passante h est le paramètre dont le choix a la plus forte influence sur les résultats.** Plus la valeur de la bande passante est élevée, plus le nombre d'observations auxquelles le noyau accorde un poids non nul est élevé. La régression locale inclura alors davantage d'observations et les résultats seront plus lissés qu'avec une faible bande passante.

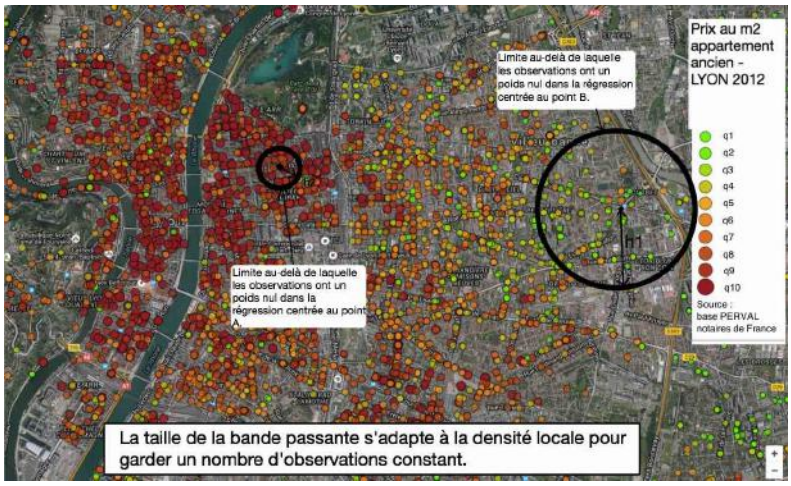


FIGURE 9.7 – Noyau adaptatif
Source : *base PERVAL*

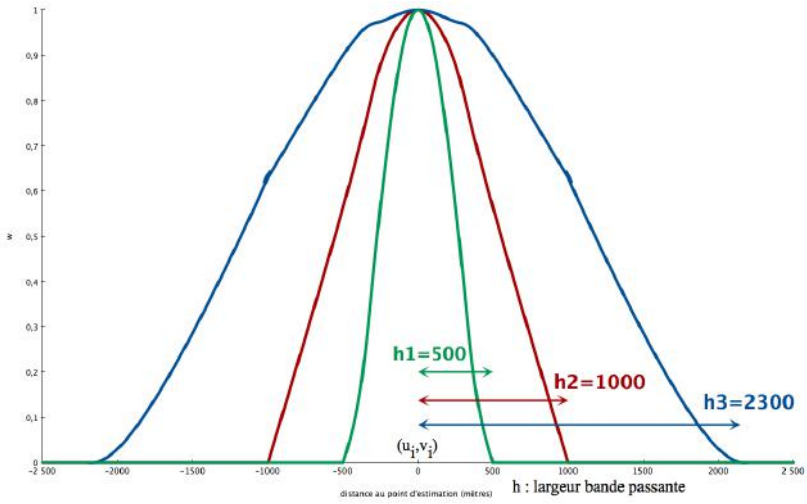


FIGURE 9.8 – Influence du choix de la bande passante sur le noyau

Lorsque la bande passante tend vers l'infini, les résultats de la régression locale s'approchent de ceux d'une régression par moindres carrés ordinaires.

Le choix de la bande passante n'est pas lié au modèle lui même, mais à la stratégie de calibration. Si le noyau inclut des points trop éloignés, la variance sera faible mais le biais élevé. Si le noyau inclut uniquement les points trop proches, le biais sera faible mais la variance élevée. Plusieurs critères statistiques peuvent aider à choisir la bande passante la plus adaptée. Le package R *GW.model* permet d'obtenir la bande passante minimisant l'un ou l'autre des deux critères : le critère de validation croisée et le critère d'Akaike corrigé (voir encadrés 9.2.1 et 9.2.1).

La valeur de la bande passante minimisant ces critères est aussi une indication précieuse quant à la pertinence d'une modélisation par une régression géographiquement pondérée. Si la bande passante tend vers le maximum possible (toute la taille de la zone d'étude, ou tous les points), alors l'hétérogénéité locale n'est probablement pas significative et la RGP n'est pas nécessaire. Inversement, une bande passante extrêmement faible doit alerter sur le risque que le processus sous-jacent soit aléatoire (GOLLINI et al. 2013). Il faut également garder en tête que la bande passante qui minimise les critères statistiques s'appuie sur la prédiction de la variable dépendante, et sur celle des coefficients de la régression (qui sont pourtant ceux utilisés ensuite pour tester la validité de l'hypothèse de non-stationarité).

Encadré 9.2.1 — Critère de validation croisée.

$$CV = \sum_{i=1}^{n} \left[y_i - \hat{y}_{\neq i}(h) \right]^2$$

$\hat{y}_{\neq i}(h)$ est la valeur de y au point i prédite lorsqu'on calibre le modèle avec toutes les observations sauf y_i. En effet, si on estimait le modèle avec l'intégralité des observations, la bande passante optimale serait 0 puisque, lorsque $h = 0$, on n'inclut aucun autre point que y_i dans la régression ; on a donc $\hat{y}_i = y_i$ ce qui est l'optimum atteignable.
La bande passante h qui minimise le score de validation croisée CV **maximise le pouvoir prédictif du modèle**.

Encadré 9.2.2 — Critère d'Akaike corrigé.

$$AIC_c(h) = 2n\ln(\hat{\sigma}) + n\ln(2\pi) + n\left\{ \frac{n + tr(S)}{n - 2 - tr(S)} \right\}$$

n est la taille de l'échantillon ; $\hat{\sigma}$ est l'estimation de la déviation standard du terme d'erreur ; $tr(S)$ est la trace de la matrice de projection (matrice chapeau) de la variable observée y sur la variable estimée $\hat{y}$.
Le critère AIC favorise un **compromis entre le pouvoir prédictif du modèle et sa complexité.** Plus la bande passante est faible, plus le modèle global est complexe. Le critère AIC favorise généralement des bandes passantes plus larges que le critère CV.

9.3 Régression géographiquement robuste

Tout comme la régression linéaire classique, la régression géographiquement pondérée est sensible aux points aberrants. Ces points distordent la surface des paramètres estimés. Puisque la régression géographiquement pondérée estime un modèle différent en chaque point de l'espace, il suffit qu'un point soit aberrant **par rapport au contexte local** pour que l'estimation soit faussée. Or il y a plus de chances pour qu'un point soit aberrant par rapport au contexte local que par rapport

au contexte global. Rechercher les points aberrants au niveau global risque donc de laisser passer des points qui sont aberrants localement mais pas globalement. Deux méthodes ont été développées pour pallier ce problème.

Méthode 1 : filtrer en fonction des résidus standardisés

L'objectif de la méthode 1 est de détecter les observations dont les résidus sont très élevés et de les exclure de la régression.

Soit $e_i = y_i - \hat{y}_i$ le résidu de l'estimation au point i. Si y_i est un point aberrant, e_i devrait avoir une valeur très élevée. Cependant, les résidus n'ont pas tous la même variance. Il faut donc les standardiser afin de pouvoir les comparer et juger de ceux qu'il est nécessaire d'éliminer de la régression.

Notons $\hat{y} = \mathbf{Sy}$ où S est la matrice chapeau définie plus haut. On a $\mathbf{e} = \mathbf{y} - \mathbf{Sy} = (\mathbf{I} - \mathbf{S})\mathbf{y}$ avec $\mathbf{e}$ le vecteur des résidus et $var(\mathbf{e}) = (\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})^{\mathbf{T}} var(\mathbf{y}) = (\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})^{\mathbf{T}}\sigma^2$ avec σ la déviation standard de y. Les variances des e_i sont donc les éléments de la diagonale de la matrice $(\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})^{\mathbf{T}}\sigma^2$, *a priori* différents.

Soit $\mathbf{Q} = (\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})^{\mathbf{T}}$ et q_{ii} le $i$ème élément de la diagonale de $\mathbf{Q}$.
$r_i = \frac{e_i}{\hat{\sigma}\sqrt{q_{ii}}}$ est appelé *résidu standardisé intérieur*.
Si le point i est aberrant, l'inclure dans l'estimation de $\hat{\sigma}^2$ risque de produire un biais. On estime donc la valeur de σ en excluant l'observation i potentiellement aberrante : σ_{-i}
$r_i^* = \frac{e_i}{\hat{\sigma_{-i}}\sqrt{q_{ii}}}$ est appelé *résidu standardisé extérieur*.
Avec la méthode 1, les observations pour lesquelles $|\mathbf{r_i^*}| > 3$ sont filtrées (le seuil de 3 est proposé par CHATFIELD 2006).

Inconvénient de la méthode : $\mathbf{Q}$ est une matrice $n * n$ dont **le temps de calcul est à ce jour rédhibitoire pour de grosses bases de données**, avec une machine dotée d'une puissance de calcul usuelle. Par exemple, BRUNSDON et al. 1996 jugent qu'au-delà de 10 000 observations l'emploi de cette méthode n'est pas envisageable.

Méthode 2 : diminuer les poids des observations aux résidus élevés

L'objectif de la méthode 2 est de diminuer le poids des observations ayant des résidus élevés (HUBER 1981). Après une première estimation du modèle, on accorde un poids $w_r(e_i)$ supplémentaire à chaque observation i. Ce poids doit être multiplié avec le poids qui varie en fonction la distance au point i. On a donc une nouvelle matrice W qui est le produit terme à terme entre l'ancienne matrice W et une matrice W_r des poids des résidus, ainsi définis :

$$w_r(e_i) = \left\{ \begin{array}{cc} 1 & si \quad |e_i| \leq 2\hat{\sigma} \\ \left[1 - (|e_i| - 2)^2\right]^2 & si \quad 2\hat{\sigma} < |e_i| < 3\hat{\sigma} \\ 0 & sinon \end{array} \right\} \tag{9.11}$$

Si aucun des résidus de la première régression n'est plus élevé que deux déviations standard, le deuxième modèle est identique au premier. Les observations dont les résidus sont compris entre deux et trois déviations standard voient leur poids diminué dans la deuxième régression, tandis que les observations dont les résidus sont supérieurs à trois déviations standard sont carrément exclues du modèle.

Discussion

La méthode 2 est beaucoup plus rapide à calculer que la méthode 1 puisque chaque cycle demande seulement le calcul des n résidus et non celui d'une matrice $n \times n$. Cependant, elle ne prend pas en compte les différences de variance entre les résidus et élimine davantage de points que la méthode 1.

Application avec R

Le package *GWmodel* permet de mettre en œuvre la régression géographiquement pondérée. La première étape consiste à calculer les distances entre toutes les observations grâce à la fonction `gw.dist`. Ensuite, la fonction `bw.gwr` permet de calculer de manière optimale, au sens d'un critère statistique donné, la bande passante de la fonction de noyau, . Enfin, les coefficients locaux de la régression géographiquement pondérée sont obtenus grâce à la fonction `gwr.robust`. Les résultats sont contenus dans un objet de classe `gwrm`, contenant en particulier un objet de type `SpatialPointsDataFrame`, dont le contenu est détaillé ci-après.

Options de la fonction `gw.dist`

— `dp.locat` : coordonnées des observations ;
— `rp.locat` : coordonnées des points sur lesquels sera calibré le modèle (par exemple : les points d'une grille régulière) ;
— `p` : régit le choix de la distance (p=1 : Manhattan ; p=2 : euclidienne) ;
— `theta` : angle avec lequel on effectue une rotation du système de coordonnées (utile pour distance Manhattan).

Options de la fonction `bw.gwr`

— `formula` : le modèle $y \sim x_1 + x_2 + ... + x_p$;
— `approach` : méthode de calcul de la bande passante optimale : CV (Validation Croisée) ou AIC (Critère d'Information d'Akaike) ;
— `kernel` : type de noyau : "gaussian", "exponential", "bisquare", "tricube", "boxcar" ;
— `adaptive` : si TRUE, la bande passante est un nombre de voisins, le noyau est adaptatif et si FALSE, la bande passante est une distance, le noyau est fixe ;
— `dMat` : la matrice de distances pré-calculée.

Options de la fonction `gwr.robust`

— `regression points` : les coordonnées géographiques des points à partir desquels le modèle sera évalué ;
— `bw` : la taille de la bande passante ;
— `filtered` : si TRUE, filtre les observations en fonction de la valeur des résidus standardisés (Méthode 1 de régression robuste). Si FALSE, estime le modèle une deuxième fois en pondérant les observations en fonction de la valeur de leurs résidus (Méthode 2 de régression robuste) ;
— `F123.test` : calcule la statistique de Fischer (défaut FALSE) ;
— `maxiter` : nombre maximum d'itérations de l'approche automatique (Méthode 2) : vaut 20 par défaut ;
— `cut1` : $\sigma cut1$ est le seuil de valeur des résidus au-delà duquel les observations ont un poids < 1 (vaut 2 par défaut) ;
— `cut2` : $\sigma cut2$ est le seuil de valeur des résidus au-delà duquel les observations ont un poids nul (vaut 3 par défaut) ;
— `delta` : seuil de tolérance de l'algorithme itératif (vaut 10^{-5} par défaut).

Interprétation des résultats : le contenu du fichier $SDF

— Le fichier $SDF est de nature "SpatialPointsDataFrame", il contient des attributs associés à des coordonnées géographiques.

— *c_x* : estimation du coefficient associé à la caractéristique x en chaque point.
— *yhat* : valeur de y prédite.
— *residual*, *Stud_residual* : résidu et résidu standardisé
— *CV_score* : score de validation croisée
— *x_SE* : erreur standard de l'estimation du coefficient devant la caractéristique x.
— *x_TV* : t-value de l'estimation du coefficient devant la caractéristique x.
— *E_weight* : poids des observations dans la régression robuste (à multiplier au poids lié à la fonction de noyau).

■ **Exemple 9.2 — Application à l'étude des prix de l'immobilier lyonnais.** La régression géographiquement pondérée permet d'étudier l'influence de la localisation d'un bien immobilier sur son prix, tout en prenant en compte l'hétérogénéité spatiale c'est-à-dire le fait que l'influence des caractéristiques d'un bien immobilier sur son prix dépende de sa localisation. Le coefficient associé à la constante de la régression géographiquement pondérée est le prix d'un appartement de référence : le prix d'un appartement, une fois prise en compte l'influence de ses caractéristiques physiques.

Signification des variables de l'exemple ci-dessous :
f_lgpx : logarithme du prix au mètre carré.
c_epoqueA : indicatrice de construction avant 1850
c_epoqueF : indicatrice de construction entre 1981 et 1991
c_epoqueG : indicatrice de construction entre 1992 et 2000
c_mmut1, 2, 3 : indicatrice d'une mutation ayant eu lieu au mois de janvier, février, mars, etc.
c_sdbn_2 : indicatrice de l'existence de deux salles de bain
c_cave1 : indicatrice de l'existence d'une cave

```
library(GWmodel)
dm.calib <- gw.dist(dp.locat=coordinates(lyon2012))

#Calcule une matrice de distances entre les points
bw0 <- bw.gwr(f_lgpx~c_epoqueG+c_mmut_1+c_mmut_2+
                c_mmut_3+c_epoqueA+c_epoqueF+c_sdbn_2+c_cave1,
              data=lyon2012, approach="AIC", kernel="bisquare",
              adaptive=TRUE,dMat=dm.calib)

gwr.robust.lyon2012 <- gwr.robust(f_lgpx~c_epoqueG+c_mmut_1+c_mmut_2+
                c_mmut_3+c_epoqueA+c_epoqueF+c_sdbn_2+c_cave1,
                bw=bw0, kernel="bisquare", filtered=FALSE, adaptive=TRUE,
                                                dMat=dm.calib)

#Extraction de la constante : prix du bien de référence (figure 9.9)

lyon2012.intercept.robust <- gwr.robust.lyon2012$SDF[,c(1)]
# 1 correspond à la position de la constante dans le fichier contenant les
    résultats de la régression.
lyon2012.intercept.robust$Intercept <- exp(lyon2012.intercept.robust$
    Intercept)

#Extraction du coefficient lié au fait d'avoir été construit avant 1850
    plutot qu'entre 1948 et 1969 (époque de référence) - figure 9.10
lyon2012.epoqueA.robust <- gwr.robust.lyon2012$SDF[,c(15)]
```

```
lyon2012.epoqueA.robust$c_epoqueA <- exp(lyon2012.intercept.robust$c_
    epoqueA)

#Estimation du modèle (non robuste) sur un carroyage de 100 mètres de coté
    (figure 9.11)
#Soit "quadrillage" un fichier de type SpatialGridDataFrame recouvrant la
    zone à étudier
dm.calib.quadrillage<- coordinates(quadrillage) gw.dist(dp.locat=
    coordinates(lyon2012),rp.locat=coordinates(quadrillage))
gwr.lyon2012<-gwr.basic(f_lgpx~c_epoqueG+c_mmut_1+c_mmut_2+c_mmut_3+c_
    epoqueA+c_epoqueF+c_sdbn_2+c_cave1,regression.point=quadrillage,bw=bw0,
     kernel="bisquare", filtered=FALSE, adaptive=TRUE,    dMat=dm.calib.
    quadrillage)
```

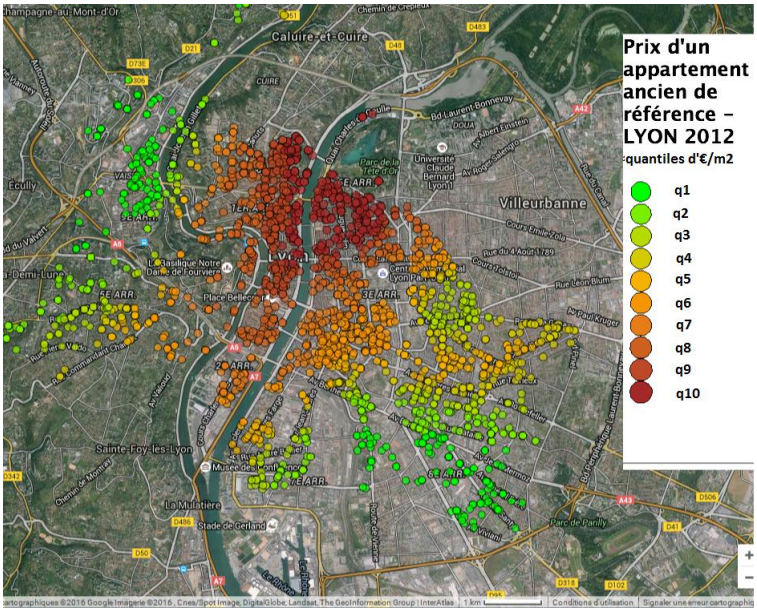


FIGURE 9.9 – Constante locale : prix du bien de référence
Source : *base PERVAL*

Coefficient	Min	1er quartile	Médiane	Moyenne	3ème quartile	Max
constante	1666	2220	2668	2705	3088	4030
époque A	0.6250	0.9533	1.1480	1.1070	1.2470	1.8190

TABLE 9.4 – Statistiques descriptives des coefficients de la RGP des prix immobiliers sur leurs caractéristiques
Source : *base PERVAL*

Les coefficients de la régression hédonique varient dans l'espace (Table 9.4). La régression géographiquement pondérée a permis de mieux appréhender la richesse spatiale de l'évolution des paramètres explicatifs des prix immobiliers puisque les estimations sont indépendantes de la frontière administrative des arrondissements. Sur les Figures 9.9 et 9.10, les points où les coefficients ont été estimés sont ceux où des transactions avaient eu lieu. Cependant, un des intérêts de la RGP est aussi de pouvoir estimer les valeurs des coefficients de façon continue. La Figure 9.11

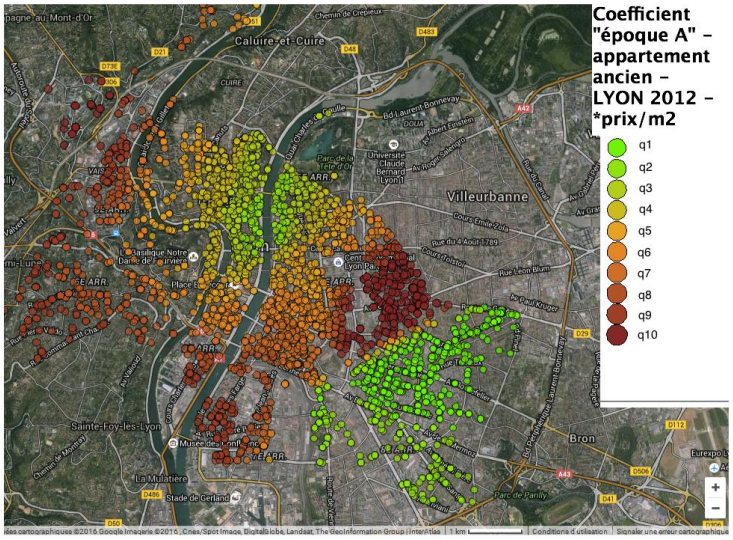


FIGURE 9.10 – Coefficient associé au fait d'avoir été construit avant 1850 plutôt qu'entre 1948 et 1969 (époque de référence)
Source : *base PERVAL*

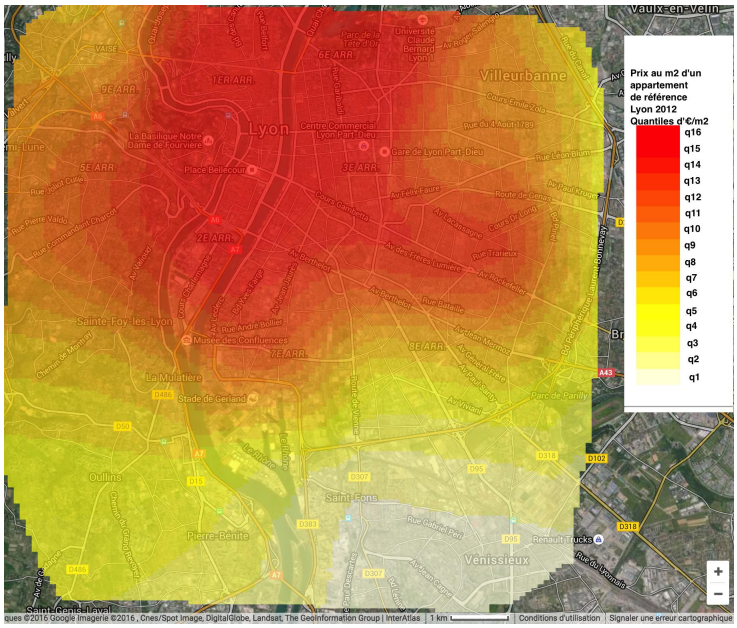


FIGURE 9.11 – Estimation des prix immobiliers sur un carroyage de 100 mètres de côté
Source : *base PERVAL*

présente une estimation des paramètres sur un quadrillage de 100 mètres de côté. La Section 9.4 présente une méthode permettant d'évaluer dans quelle mesure la variation spatiale des paramètres est significative. ■

9.4 Qualité des estimations

9.4.1 Précision de l'estimation des coefficients

Quand on estime une RGP avec un noyau adaptatif dans une zone où les observations sont peu denses, les points servant à calibrer le modèle peuvent tous avoir un poids très faible (ils sont situés à une grande distance du point d'estimation).

Soit $\mathbf{C}$ la matrice telle que :

$$\hat{\beta}(u_i, v_i) = (\mathbf{X^T W_{(u_i,v_i)} X})^{-1} \mathbf{X^T W_{(u_i,v_i)} y} = \mathbf{CY} \tag{9.12}$$

La variance du paramètre estimé est :

$$Var\left[\hat{\beta}(u_i, v_i)\right] = \mathbf{CC^T}\sigma^2 \tag{9.13}$$

Avec σ^2 la somme des résidus normalisés de la régression locale :

$$\sigma^2 = \sum_i (y_i - \hat{y}_i)/(n - 2\nu_1 + \nu_2) \tag{9.14}$$

$$\nu_1 = tr(\mathbf{S}) \tag{9.15}$$

$$\nu_2 = tr(\mathbf{S^T S}) \tag{9.16}$$

$$\hat{Y} = \mathbf{SY} \tag{9.17}$$

Une fois que la variance de chaque paramètre a été estimée, les erreurs standard sont obtenues avec l'équation 9.18

$$SE(\hat{\beta}(u_i, v_i)) = \sqrt{Var}\left[\hat{\beta}(u_i, v_i)\right] \tag{9.18}$$

On peut ainsi calculer des intervalles de confiance pour les coefficients.

Application avec R

Le fichier $SDF contenant les résultats de la régression géographiquement pondérée permet d'accéder aux erreurs standard associées aux différents coefficients. Ainsi, dans le cas de l'exemple des prix de l'immobilier lyonnais développé précédemment :

- y : prix de vente ;
- `yhat` : prix de vente estimé ;
- `Intercept_SE` : erreur standard du coefficient associé à la constante ;
- `Intercept_TV` : taux de variation du coefficient associé à la constante.

9.4.2 Test de la non-stationnarité des coefficients

La RGP relâche l'hypothèse que les coefficients sont stationnaires dans une certaine zone géographique. Pour s'assurer de la pertinence du modèle, il est intéressant de tester la non-stationnarité des coefficients. **Les coefficients varient-ils suffisamment dans l'espace pour rejeter l'hypothèse qu'ils sont constants sur toute la surface de l'étude ?**

En termes statistiques, la question équivaut à :
— $H_0 : \forall k, \beta_k(u_1, v_1) = \beta_k(u_2, v_2) = ... = \beta_k(u_n, v_n)$
— $H_1 : \exists k$, tous les $\beta_k(u_i, v_i)$ ne soient pas égaux.
Pour répondre à cette question, on peut utiliser une méthode de simulation du type "simulation de Monte Carlo".

Principe : S'il n'y avait pas de phénomène spatial sous-jacent, on pourrait permuter les coordonnées géographiques des observations aléatoirement dans l'espace et la variance resterait inchangée. Lors d'une simulation de Monte Carlo, on permute n fois les coordonnées géographiques des observations. On obtient donc n estimations de la variance spatiale des coefficients. On peut ensuite estimer la $p - value$ de la variabilité spatiale des coefficients et rejeter - ou non - l'hypothèse nulle selon laquelle ils sont stables dans l'espace.
Rappelons néanmoins que les méthodes simulant une distribution par permutation spatiale des observations dépendent du jeu de données initial. LEUNG et al. 2000 décrivent une méthode plus robuste et moins gourmande en temps de calcul pour tester la non-stationnarité des coefficients.

Application avec R

Fonction `montecarlo.gwr`
— mêmes paramètres que la fonction `gwr.robust` ;
— `nsims` : nombre de simulations ;
— `sortie` : vecteur contenant les p-values de tous les paramètres de la RGP.

9.5 Une application prédictive

La régression géographiquement pondérée a surtout été utilisée pour mettre en évidence l'hétérogénéité spatiale. Comme les autres méthodes de régression, elle peut aussi être utilisée à des fins prédictives, par exemple pour imputer des valeurs à des unités non échantillonnées dans le cadre d'un sondage. Cette partie de l'article s'appuie sur un travail réalisé par E.Lesage et J-M. Floch pour les JMS de 2015 [1], et également présenté aux Journées des Méthodes Avancées pour l'Analyse de Sondages Complexes de 2016. Dans les méthodes d'estimation sur petits domaines, on utilise de plus en plus fréquemment une approche basée sur des modèles qui utilisent des estimateurs BLUP, Best linear unbiased Predictors (CHAMBERS et al. 2012). Les valeurs des unités non échantillonnées sont remplacées par les valeurs prédites à partir d'un modèle dont les paramètres sont estimés à l'aide des valeurs des unités échantillonnées. Une extension de ces méthodes a été proposée par CHANDRA et al. 2012 dans un cadre non stationnaire, en recourant à la régression géographiquement pondérée. L'utilisation de la régression géographiquement pondérée dans les méthodes d'estimation sur petits domaines semble être préférée dans la littérature récente aux méthodes issues de l'économétrie spatiale recourant notamment aux modèles spatiaux autoregressifs (SAR). La régression géographiquement pondérée offre une façon plus flexible de prendre en compte la variabilité spatiale des phénomènes. Cette prise en compte de l'hétérogénéite spatiale doit théoriquement améliorer la précision des estimateurs.

1. Journées de Méthodologie Statistique organisées par l'Insee tous les trois ans environ. Le diaporama de la présentation est accessible à l'adresse `https://maasc2016.sciencesconf.org/data/pages/7.DiapoFlochJeanMichel.pdf`

9.5.1 Présentation du problème

À l'Insee, des travaux empiriques ont utilisé la RGP pour construire des estimateurs à partir de données issues du recensement de la population concernant les quartiers prioritaires. Dans ces quartiers, 40 % des logements sont enquêtés (sur une période de cinq ans), mais le plan de sondage n'est pas optimal, l'appartenance à un quartier prioritaire ne faisant pas partie des variables d'équilibrage. La demande d'une information précise sur ces quartiers étant forte, on a cherché à mobiliser des sources administratives exhaustives ou quasi-exhaustives (Données fiscales, données de l'assurance maladie) pour améliorer la précision des estimateurs. Pour ce faire, on estimait sur les logements de l'échantillon du recensement de la population (RP) un modèle dans lequel la variable d'intérêt était une variable du recensement, les variables auxiliaires des variables tirées des sources administratives, bien corrélées à la variable d'intérêt. Les estimateurs permettaient de prédire une valeur sur les logements non échantillonnés. **L'estimateur du total de la variable d'intérêt correspond à la somme des valeurs observées pour les unités échantillonnées et des valeurs prédites pour les unités non échantillonnées, les poids de sondage n'intervenant plus dans ce calcul**.

Ces travaux empiriques reposaient sur une modélisation utilisant la régression géographiquement pondérée afin de tenir compte de la forte hétérogénéité constatée dans les données urbaines. Mais le gain de précision, par rapport à un modèle non spatial, n'avait pas été étudié. C'est pourquoi on propose une comparaison de trois estimateurs, à partir d'un dispositif expérimental reposant sur des données réelles administratives composées de la source Filosofi, qui permet de calculer la population des ménages à bas revenus, de la source CNAM (Caisse Nationale d'Assurance Maladie) qui fournit le nombre de bénéficiaires et de la CMUC (Couverture Maladie Universelle Complémentaire). La source Filosofi est quasi-exhaustive et permet de disposer des "vraies" valeurs du nombre d'individus ayant des revenus inférieurs au taux de pauvreté.

Les deux sources sont localisées et peuvent être théoriquement appariées à partir de leurs coordonnées géographiques. Pour des raisons de confidentialité, il n'a pas été possible de le faire, et on a calculé le nombre de personnes à bas revenus et le nombre de bénéficiaires de la CMUC sur un maillage formé de carreaux de 100 m de côté, compromis jugé acceptable avec l'utilisation de données individuelles. Ces carreaux de 100 m jouent le rôle d'individus statistiques sur lesquels on va effectuer les mesures.

Le territoire d'intérêt est la commune de Rennes. On tire dans la base de données un échantillon de 40 % des carreaux, comme ce qui est fait dans le recensement de la population. Ces carreaux vont servir de support à l'estimation du nombre de personnes à bas revenus (figure 9.12). On dispose de toutes les informations, mais pour le modèle, les bas revenus ne sont connus que pour les carreaux de l'échantillon, tandis que les bénéficiaires de la CMUC le sont pour l'ensemble des carreaux. On sélectionne un échantillon de taille $n = 856$, qu'on nomme s, par tirage aléatoire simple sans remise (moins complexe que le tirage opéré pour le RP). Le taux de sondage est $n/N = 40$ %. De plus, on note r le complémentaire de l'échantillon s dans U (l'ensemble des carreaux habités sur le territoire de Rennes). Les calculs effectués au niveau du carreau permettent des calculs sur la maille Iris, chaque carreau étant affecté à un Iris (l'Iris est le plus petit découpage administratif français).

Il existe une liaison linéaire forte entre le nombre de personnes à bas revenu et le nombre de bénéficiaires de la CMUC. Les ordonnées à l'origine varient peu d'un carreau à l'autre. La valeur des pentes varient sensiblement de 1.6 à 3.3. Le gradient des situations locales est représenté sur la figure 9.13. Dans l'approche dite "basée sur le modèle", on prédit les valeurs y_i des carreaux non échantillonnés grâce au modèle estimé à partir de l'ensemble des données de l'échantillon et de l'information auxiliaire x disponible pour les carreaux non échantillonnés. On construit trois estimateurs, j désignant l'Iris :

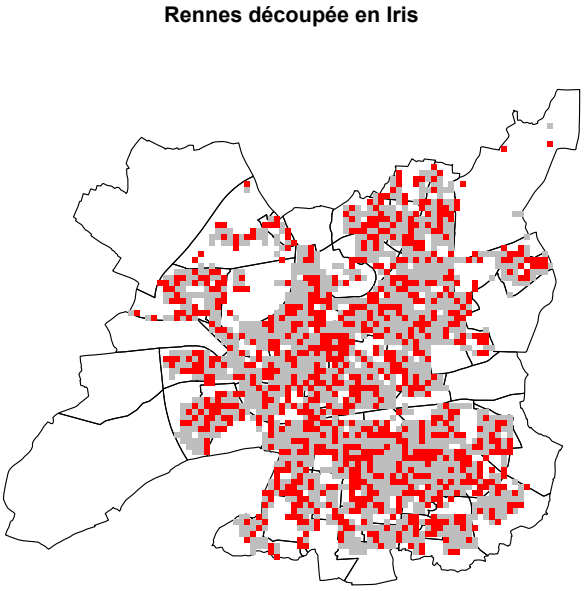


FIGURE 9.12 – Carreaux de 100 m habités à Rennes, échantillonnés (en rouge) ou non (en gris)

Définition 9.5.1 — L'estimateur de Horvitz-Thompson.

$$\widehat{t}_y(j) = \frac{N}{n} \sum_{i \in s_j} y_i \tag{9.19}$$

Définition 9.5.2 — L'estimateur basé sur la régression "classique", sans prise en compte de l'hétérogénéité spatiale.

$$\widehat{t}_{y,reg}(j) = \sum_{i \in s_j} y_i + \sum_{i \in r_l} \tilde{y}_l \tag{9.20}$$

où $\tilde{y}_l = \beta^T x_l$

Définition 9.5.3 — L'estimateur basé sur la régression géographique pondérée.

$$\widehat{t}_{y,RGP}(j) = \sum_{i \in s_j} y_i + \sum_{i \in r_l} \breve{y}_l \tag{9.21}$$

où $\breve{y}_l = \hat{\beta}_l^T x_l$ et $\widehat{\beta}_l$ est le vecteur des coefficients de la régression géographique pondérée pour le carreau l.

9.5.2 Résultats

On répète $K = 1000$ fois ce processus. On obtient pour chaque Iris 1 000 valeurs pour chacun des trois estimateurs. À partir de ces 1 000 valeurs, on construit des estimations Monte Carlo des biais et des erreurs quadratiques moyennes des estimateurs.

Si on note $\widehat{t}_y(j)^{(k)}$ l'estimateur du total de la variable y pour l'Iris j et pour la simulation k, on

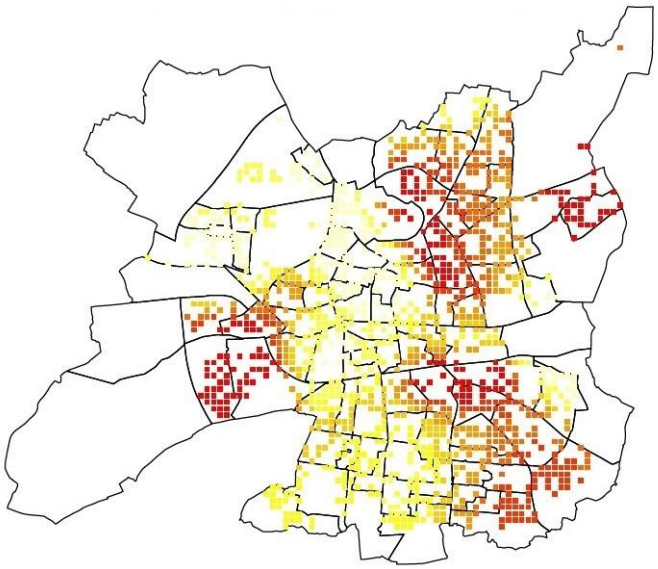


FIGURE 9.13 – Représentation graphique des pentes des régressions géographiques pondérées
Note : L'échelle utilisée est une "échelle de chaleur" qui va de la couleur jaune (valeurs les plus fortes) à la couleur rouge (valeurs les plus faibles).

peut calculer l'erreur quadratique moyenne "Monte Carlo" défini par :

$$EQM(\hat{t}_y(j)) = K^{-1} \sum_{k=1}^{K} (\hat{t}_y(j)^{(k)} - t_y(j))^2 \tag{9.22}$$

puisque l'on connait le total exact $t_y(j)$.

On en déduit l'indicateur qui va servir à comparer les résultats des trois estimations, la racine carrée de l'erreur quadratique moyenne relative :

$$RCEQMR(\hat{t}_y(j)) = \frac{\sqrt{EQM(\hat{t}_y(j))}}{t_y(j)} \tag{9.23}$$

Les Iris de la commune de Rennes sont classés par ordre de taille de population croissante, et on représente sur la figure 9.14 les RCEQMR pour chacun des Iris.

Le premier résultat est l'amélioration de la précision dans les deux approches par les modèles de régression, du fait de la bonne relation linéaire entre la variable y (les personnes à bas revenus) et la variable x (les bénéficiaires de la CMUC). La RCEQMR est de l'ordre de 0.4 pour l'estimateur de Horwitz-Thompson, de l'ordre de 0.12 pour les modèles de régression. La différence entre la régression et la RGP n'est pas très visible sur le graphique de la figure 9.14. Les résultats sont très proches. Les box-plot de la figure 9.15 permettent d'aller un peu plus loin dans la comparaison.

Au vu de la figure 9.14 l'estimateur RGP s'avère néanmoins meilleur que l'estimateur par la régression : pour 75 % des IRIS, la RCEQMR de l'estimateur RGP est inférieure à 0.156, la valeur correspondante pour l'estimateur par la régression étant de 0.178.

9.6 Précautions particulières

9.6.1 Multicolinéarité et corrélation entre les coefficients

Détecter la colinéarité

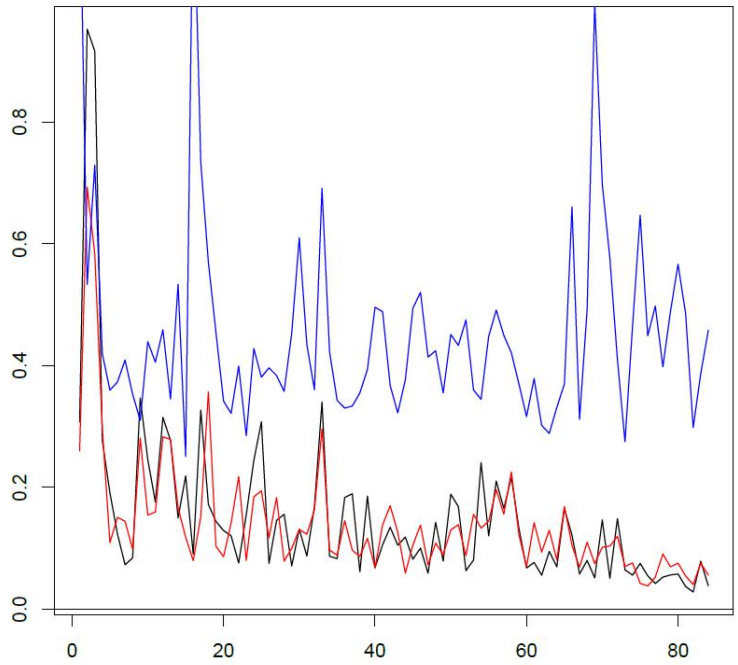


FIGURE 9.14 – RCEQMR (RRMSE sur la figure) de l'estimateur de Horwitz-Thompson (en bleu), de l'estimateur par la régression (en noir) et de l'estimateur par la RGP (en rouge), selon les Iris, classés par taille croissante.

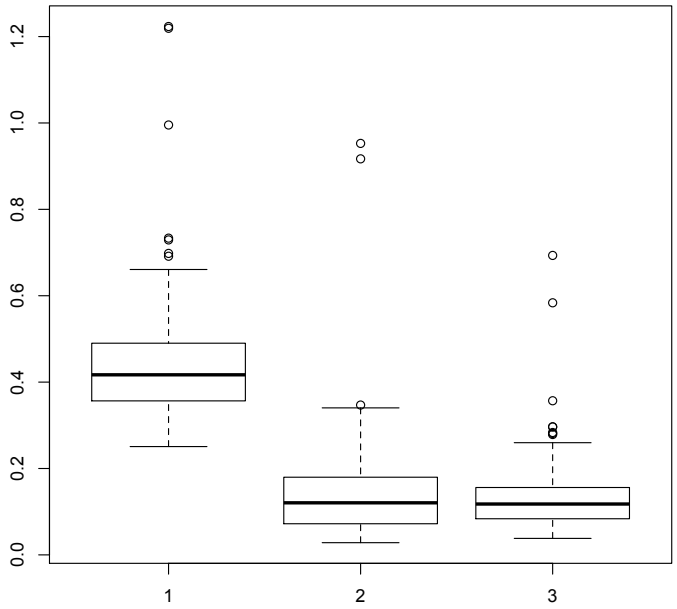


FIGURE 9.15 – Box-plot des RCEQMR de l'estimateur de Horwitz-Thompson (1), de l'estimateur par la régression (2) et de l'estimateur par la RGP (3)

Pour pouvoir estimer les très nombreux coefficients d'une régression géographiquement pondérée, la technique des moindres carrés pondérés impose de nombreuses contraintes sur les paramètres de la régression (LEUNG et al. 2000). Ces contraintes peuvent relier les coefficients de la RGP et créer des problèmes de multicolinéarité.

La multicolinéarité entre les variables peut être responsable d'une grande instabilité dans les coefficients (changement de signe lors de l'ajout d'une nouvelle variable dans la régression) ; du signe contre-intuitif de l'un des coefficients de la régression, ou encore d'erreurs standard des paramètres élevées (WHEELER et al. 2005). Si la structure de corrélation des données est hétérogène dans l'espace, certaines régions peuvent présenter une colinéarité entre leurs variables, tandis que d'autres n'en présenteront pas.

La fonction `gwr.collin.diagno` du package *GWmodel* permet de mettre en œuvre plusieurs types de détection de la colinéarité, notamment les corrélations locales entre les paires de coefficients et les facteurs d'inflation de la variance (VIF) pour chaque coefficient. Ces éléments sont détaillés dans GOLLINI et al. 2013 où des exemples d'application avec R sont présentés.

Encadré 9.6.1 — Facteur d'inflation de la variance : VIF. Soit R_j^2 le coefficient de détermination de la régression de la variable X_j avec les $p-1$ autres variables.

$$VIF_j = \frac{1}{1-R_j^2}$$

Si R_j^2 tend vers 1, VIF_j tend vers $+\infty$ d'où le terme "inflation de la variance". En général, la littérature considère qu'il existe un problème de multicolinéarité lorsqu'un VIF est supérieur à 10, ou que la moyenne des VIF est supérieure à 2 (CHATTERJEE et al. 2015).

Prendre en compte la colinéarité

Une méthode permettant de réduire les problèmes de colinéarité implémentée dans le package *GWModel* est la régression ridge. Le principe est d'augmenter le poids des éléments diagonaux de la matrice de variances-covariances pour diminuer le poids des éléments hors-diagonale (qui contiennent les termes de colinéarité). Dans le cas général, on peut écrire :

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y \tag{9.24}$$

L'inconvénient de cette méthode est que $\hat{\beta}$ est biaisé et que les erreurs standard ne sont plus disponibles.
Dans le cas de la régression géographiquement pondérée, on peut définir une régression ridge locale telle que :

$$\hat{\beta}(u_i, v_i) = (X^T W(u_i, v_i) X + \lambda I(u_i, v_i))^{-1} X^T W(u_i, v_i) Y \tag{9.25}$$

$\lambda I(u_i, v_i)$ est la valeur de λ à la localisation (u_i, v_i). Il est également possible d'utiliser un critère statistique tel que le score de validation croisée pour choisir la bande passante de la régression locale ridge.

9.6.2 Interprétation des paramètres

Problème du test des hypothèses multiples

Lors de l'estimation d'une régression géographiquement pondérée, on obtient en chaque point une évaluation de la significativité de chaque coefficient grâce aux t-values calculées. Pour chaque coefficient, on obtient donc autant de t-values que de points où elles ont été estimées. On se heurte alors au problème du test des hypothèses multiples présenté au chapitre 3 dans le cas des indicateurs

locaux d'autocorrélation spatiale.
Si l'on estime la significativité d'un coefficient en 100 localisations avec un seuil de significativité défini à 95 %, on s'attend à juger le coefficient significatif au moins en 5 localisations, et ceci simplement en raison du principe statistique du test, indépendamment de toute corrélation réelle entre la variable dépendante et les variables explicatives. Pour pallier ce problème, on peut utiliser une méthode d'ajustement de Bonferroni qui augmente la valeur du seuil au-delà duquel le résultat du test local sera jugé non significatif - à niveau de significativité global constant. Cependant les méthodes d'ajustement ont l'inconvénient d'être souvent trop restrictives ce qui peut conduire à juger certains coefficients non significatifs alors qu'ils le sont en réalité.
BRUNSDON et al. 1998 conseillent d'utiliser les t-values produites lors de l'estimation d'une RGP avec précaution. Ils considèrent qu'une surface avec une proportion importante de coefficients très divers localement est un meilleur indicateur de non-stationnarité locale qu'une surface où seule une faible proportion de coefficients dépassent une valeur significative.

Effet du contexte local ou mauvaise spécification

Avant d'interpréter les valeurs des coefficient locaux comme des caractéristiques du contexte local, il est important d'explorer la possibilité d'une mauvaise spécification du modèle. Par exemple, le fait que l'influence d'avoir un garage sur le prix d'un bien immobilier dépende de la localisation peut être dû au fait que la densité de parkings publics varie dans l'espace, ou bien que le modèle hédonique est mal spécifié.

Interprétation de la constante locale

Dans une régression géographiquement pondérée, la constante peut varier localement. Il y a donc un risque qu'elle capture tout le pouvoir explicatif des variables exogènes, en particulier lorsque celles-ci ont une influence nettement plus marquée en certaines localisations (phénomène de clustering spatial). Dans ce cas, les variables explicatives sembleront non significatives. Si l'on soupçonne un tel phénomène, on peut utiliser une régression géographiquement pondérée dite "mixte" dans laquelle la constante ne varie pas.

Conclusion

Proposée en 1998 par BRUNSDON et al. 1998, la RGP a fait l'objet de nombreuses applications pratiques dans les études géographiques et épidémiologiques notamment. Les fondements théoriques ont été considérablement approfondis. Si quelques auteurs ont mis en évidence certaines limites de la méthode, notamment les problèmes de colinéarité (WHEELER et al. 2005, GRIFFITH 2008), elle est désormais partie intégrante des outils d'analyse spatiale. Elle est présentée dans les ouvrages généraux (WALLER et al. 2004, SCHABENBERGER et al. 2017, LLOYD 2010, FISCHER et al. 2009) mais aussi dans les manuels d'économétrie spatiale(ARBIA 2014). Des extensions de la méthode (modèles linéaires généralisés) ont également été proposées.

La RGP peut être utilisée de deux façons différentes. D'un côté elle peut servir de méthode exploratoire pour détecter les zones où apparaissent des phénomènes spatiaux particuliers et les soumettre à une étude approfondie. D'un autre côté, elle peut aider à la construction d'un modèle pertinent : la détection d'une non-stationnarité spatiale est alors symptomatique d'un problème dans la définition du modèle global. BRUNSDON et al. 1998 estiment que la plupart des assertions faites à un niveau global sur la relation spatiale entre les objets mériteraient d'être examinées au niveau local à l'aide de la RGP pour tester leur validité.
La dépendance spatiale entre les termes d'erreur diminue lorsqu'une RGP est utilisée puisque l'autocorrélation spatiale est parfois le résultat d'une instabilité des paramètres non modélisée

(LE GALLO 2004). De plus, la RGP permet de calculer des indicateurs d'autocorrélation spatiale sur une variable, conditionnellement à la distribution spatiale des autres variables, ce qui n'est pas possible avec les indicateurs d'autocorrélation spatiale univariés présentés dans le chapitre 3. Nous encourageons donc à étudier conjointement la dépendance spatiale - avec les indicateurs d'autocorrélation spatiale - et l'hétérogénéité spatiale - avec la régression géographiquement pondérée.

Références - Chapitre 9

ARBIA, Giuseppe (2014). *A primer for spatial econometrics : with applications in R*. Springer.

BRUNSDON, Chris, A Stewart FOTHERINGHAM et Martin E CHARLTON (1996). « Geographically weighted regression : a method for exploring spatial nonstationarity ». *Geographical analysis* 28.4, p. 281–298.

BRUNSDON, Chris, Stewart FOTHERINGHAM et Martin CHARLTON (1998). « Geographically weighted regression ». *Journal of the Royal Statistical Society : Series D (The Statistician)* 47.3, p. 431–443.

CHAMBERS, Ray et Robert CLARK (2012). *An introduction to model-based survey sampling with applications*. T. 37. OUP Oxford.

CHANDRA, Hukum, Ray CHAMBERS et Nicola SALVATI (2012). « Small area estimation of proportions in business surveys ». *Journal of Statistical Computation and Simulation* 82.6, p. 783–795.

CHATFIELD, Chris (2006). « Model uncertainty ». *Encyclopedia of Environmetrics*.

CHATTERJEE, Samprit et Ali S HADI (2015). *Regression analysis by example*. John Wiley & Sons.

FISCHER, Manfred M et Arthur GETIS (2009). *Handbook of applied spatial analysis : software tools, methods and applications*. Springer Science & Business Media.

GOLLINI, Isabella et al. (2013). « GWmodel : an R package for exploring spatial heterogeneity using geographically weighted models ». *arXiv preprint arXiv :1306.0413*.

GRIFFITH, Daniel A (2008). « Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR) ». *Environment and Planning A* 40.11, p. 2751–2769.

HUBER, Peter (1981). « J. 1981. Robust Statistics ». *New York : John Wiley*.

LE GALLO, Julie (2004). « Hétérogénéité spatiale ». *Économie & prévision* 1, p. 151–172.

LEUNG, Yee, Chang-Lin MEI et Wen-Xiu ZHANG (2000). « Statistical tests for spatial nonstationarity based on the geographically weighted regression model ». *Environment and Planning A* 32.1, p. 9–32.

LLOYD, Christopher D (2010). *Local models for spatial analysis*. CRC press.

SCHABENBERGER, Oliver et Carol A GOTWAY (2017). *Statistical methods for spatial data analysis*. CRC press.

WALLER, Lance A et Carol A GOTWAY (2004). *Applied spatial statistics for public health data*. T. 368. John Wiley & Sons.

WHEELER, David et Michael TIEFELSDORF (2005). « Multicollinearity and correlation among local regression coefficients in geographically weighted regression ». *Journal of Geographical Systems* 7.2, p. 161–187.

10. Échantillonnage spatial

CYRIL FAVRE-MARTINOZ, MAËLLE FONTAINE, RONAN LE GLEUT, VINCENT LOONIS
Insee

Résumé

Dans ce chapitre, nous nous intéressons à l'utilisation de l'information géographique dans le contexte de l'échantillonnage. Cette information peut être utilisée à différents moments dans le processus de conception d'un plan de sondage. Dans une grande partie des enquêtes en face-à-face, des plans de sondages à plusieurs degrés sont utilisés afin de réduire les coûts de collecte en concentrant géographiquement les interviews. Un géoréférencement fin des unités statistiques à échantillonner s'avère déterminant pour la constitution des entités à sélectionner au(x) premier(s) degré(s). Cette information géographique peut également être mobilisée au moment de la sélection de l'échantillon, afin d'améliorer l'efficacité statistique de celui-ci en présence d'autocorrélation spatiale positive des variables d'intérêt.

R La lecture préalable des chapitres 2 : "Codifier la structure de voisinage" et 3 : "Indices d'autocorrélation spatiale" est recommandée.

Introduction

Le projet Geostat 2 d'Eurostat (2015-2017) visait à fournir un cadre de référence permettant une production efficiente et une utilisation aisée d'information statistique finement localisée. Concernant les enquêtes par sondage, le rapport final du projet : *A Point-based Foundation for Statistics*, identifie au moins trois phases de la conception d'une enquête qui pourraient bénéficier d'une base de sondage géocodée. Premièrement, en amont, lorsque le mode de collecte est le face-à-face, la connaissance précise de la localisation de l'ensemble des unités statistiques permet la création d'unités primaires [1] (UP). La connaissance des caractéristiques de ces UP facilite la gestion du réseau d'enquêteurs tout en maintenant les qualités statistiques de l'échantillonnage. Deuxièmement, quel que soit le mode de collecte, l'information géographique permet, sous certaines conditions, d'améliorer la précision des estimations en mobilisant des méthodes d'échantillonnage spatial. Troisièmement, lors de la phase de collecte, la connaissance de la localisation des unités statistiques échantillonnées permet de faciliter leur repérage quand la qualité de l'adressage n'est pas suffisante.

Dans ce chapitre nous nous intéressons exclusivement aux deux premiers points. Nous rappelons brièvement en première partie le cadre de la théorie des sondages. Ensuite, la deuxième partie présente une méthode de constitution d'unités primaires ayant un nombre constant d'unités statistiques, tout en ayant une faible étendue géographique. La présentation des différentes méthodes d'échantillonnage spatial constitue la troisième partie, alors que la dernière partie compare leurs propriétés de façon empirique par simulation.

Parmi la riche littérature existant sur le sujet, nous nous appuyons sur, ou orientons le lecteur vers BENEDETTI et al. 2015.

10.1 Généralités

L'objectif de la théorie des sondages est d'estimer la valeur d'un paramètre θ mesuré sur une population U de taille N [2]. Ce paramètre est fonction des valeurs prises par une, ou plusieurs, variable(s) d'intérêt associée(s) à chacun des individus de la population. On note y_i la valeur de la variable d'intérêt y pour l'individu i de U. Le sondeur n'a accès aux y_i que pour une sous-partie de la population appelée échantillon et notée s. Il agrège les valeurs observées sur l'échantillon par une fonction, appelée estimateur, prenant la valeur $\hat{\theta}(s)$ pour s. Le passage de s à θ, grâce à $\hat{\theta}(s)$, est l'inférence statistique, dont les propriétés ne sont connues que si le choix de s est aléatoire.

Un plan de sondage est une loi de probabilité sur l'ensemble $\mathscr{P}(U)$ des parties (échantillons) de U. La notation classique d'un échantillon aléatoire à valeur dans $\mathscr{P}(U)$ est $\mathbb{S}$. Un plan de sondage pour lequel tous les échantillons de taille différente de n ($n \in \mathbb{N}^*$) ont une probabilité nulle d'être sélectionnés est dit de taille fixe n. La manipulation d'une loi de probabilité sur $\mathscr{P}(U)$ est généralement complexe. C'est pourquoi le statisticien d'enquête travaille avec des résumés de la loi de $\mathbb{S}$: les probabilités d'inclusion simple et double. Elles font respectivement référence aux probabilités qu'un individu donné ou qu'un couple donné d'individus soit sélectionné : $\pi_i = \mathbb{P}(i \in \mathbb{S})$ et $\pi_{ij} = \mathbb{P}((i,j) \in \mathbb{S})$.

1. Les unités primaires constituent une partition de la population selon des critères géographiques. La sélection dans un premier temps d'UP puis d'individus dans ces UP est de nature à concentrer la collecte et à réduire les coûts lorsque l'enquête se déroule en face-à-face.

2. Dans ce chapitre, contrairement aux chapitres précédents, la notion de "taille" d'une zone géographique renvoie au nombre d'entités présentes à l'intérieur, et non pas à sa surface.

L'estimation de θ par $\hat{\theta}$ est entachée d'erreurs multiples :
— erreur d'échantillonnage
— erreur de couverture : existence d'individus ne pouvant jamais être sélectionnés ;
— erreur de non-réponse : existence d'individus pour lesquels la valeur de y_i est inconnue alors même qu'ils sont échantillonnés ;
— erreur de mesure : fait de récolter la valeur y_i^* au lieu de y_i.

Un estimateur dont l'espérance est différente de θ est biaisé, alors que la variabilité des différentes valeurs $\hat{\theta}(s)$ est appréhendée par la variance de $\hat{\theta}$. L'objectif est de rendre le biais et la variance aussi petits que possible, en prêtant une attention particulière aux conditions de collecte de l'information et/ou en choisissant judicieusement le plan de sondage.

Parmi les différents paramètres à estimer, le plus classique est le total d'une variable d'intérêt : $\theta = t_y = \sum_{i \in U} y_i$. Parmi les différents estimateurs possibles de t_y, on s'intéresse à l'estimateur de Narain-Horvitz-Thompson : $\hat{t}_y = \sum_{i \in \mathbb{S}} y_i/\pi_i$. En l'absence d'erreurs de couverture, de non-réponse, et de mesure, cet estimateur est sans biais. Sa variance pour un plan de taille fixe est :

$$V(\hat{t}_y) = -\frac{1}{2} \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \tag{10.1}$$

L'analyse de l'équation 10.1 donne des indications quant aux plans de sondage à retenir pour estimer de manière précise la quantité t_y. Si les π_i sont proportionnelles aux y_i, la variance est nulle. Par rapport à cet idéal inatteignable, l'alternative de second rang consisterait à retenir, dans le cas d'une enquête, des π_i proportionnelles à x_i, où x est une variable auxiliaire connue pour tous les individus de U et qui est corrélée à y.

Ce résultat est valable quand l'enquête est mono-thème (une seule variable d'intérêt y). L'utilisation de telles probabilités pour une autre variable d'intérêt y' non corrélée à x peut en effet conduire à des estimations très imprécises. C'est pourquoi, quand l'enquête est multi-thèmes, le statisticien préfère souvent choisir des probabilités d'inclusion simple constantes. Elle permet "de rendre minimales les variances que l'on obtiendrait dans les configurations les plus défavorables (on parle d'optique MINIMAX), [...] c'est-à-dire pour les variables qui sont le plus de nature à détériorer la précision des estimations" (ARDILLY 2006). À probabilités d'inclusion simple fixées, il est souhaitable que le plan attribue des π_{ij} grands lorsque l'écart entre y_i/π_i et y_j/π_j est grand, et vice versa. Dans le cas de variables spatialisées, si on suppose que deux individus proches géographiquement ont des valeurs proches de y_i/π_i et y_j/π_j, il conviendra, à probabilités d'inclusion fixées, de privilégier la sélection d'individus éloignés plutôt que proches.

10.2 Constituer des unités primaires de faible étendue et de taille constante

10.2.1 Pourquoi ?

Quand les contraintes organisationnelles se traduisent par un mode de collecte en face-à-face sur un territoire dont la densité de population est faible, la méthode généralement retenue est celle du sondage à deux degrés. Afin de réduire les coûts liés aux déplacements des enquêteurs, le premier degré conduit à la sélection d'entités géographiques (unités primaires, UP) dont l'étendue géographique doit être la plus petite possible. De manière simplifiée, une UP ainsi sélectionnée est affectée ensuite à un seul enquêteur. Au sein de chaque UP sont sélectionnées des unités secondaires (US) correspondant aux unités statistiques que l'on souhaite interroger (individus dans des résidences principales, entreprises). Afin d'assurer une charge de travail suffisante à l'enquêteur sur une ou plusieurs enquêtes, chaque UP doit par ailleurs comporter un nombre minimal d'unités secondaires.

Pour un réseau constitué de m enquêteurs et un échantillon final de n unités secondaires, m UP sont sélectionnées proportionnellement à leur nombre d'unités secondaires : $\pi_i^{(1)} = m(N_i/N)$ pour l'UP i regroupant un total de N_i unités secondaires. En admettant que m divise n, dans chacune de ces m UP, n/m US sont tirées selon un plan à probabilités égales : $\pi_j^{i(2)} = n/(mN_i)$ pour l'unité secondaire j de l'UP i. La probabilité d'inclusion finale est constante : $\pi_j^i = \pi_i^{(1)}\pi_j^{i(2)} = n/N$.

Quand la maille géographique la plus fine disponible dans la base de sondage reste grossière, par exemple communale, les UP sont constituées par regroupement de ces mailles. Il est plus délicat d'en maîtriser la taille. Le plan ne bénéficie pas, au premier degré de tirage, de la propriété MINIMAX évoquée précédemment puisque les probabilités sont proportionnelles à la taille. Il est possible de retrouver cette propriété si les UP sont de taille constante. La constance de la taille des UP peut par ailleurs être souhaitable pour d'autres caractéristiques liées à la coordination des échantillons (sélection d'échantillons disjoints ou d'échantillons inclus les uns dans les autres). On note que :

— le complémentaire $\overline{\mathbb{S}} = U \backslash \mathbb{S}$ d'un échantillon aléatoire $\mathbb{S}$ tiré à probabilités égales, est lui-même à probabilités égales ;
— un échantillon aléatoire $\mathbb{S}_2$, sélectionné à probabilités constantes dans un échantillon $\mathbb{S}_1$ lui-même sélectionné à probabilités constantes, est à probabilités constantes.

L'idéal est donc de construire des UP de faible étendue géographique et ayant le même nombre d'unités secondaires.

10.2.2 Comment ?

Le problème de constitution d'unités primaires de faible étendue géographique et de taille constante est un cas particulier du problème plus général de classification sous contraintes de taille, qui a connu un intérêt renouvelé dans la littérature récente (MALINEN et al. 2014, GANGANATH et al. 2014, TAI et al. 2017). Il s'agit en effet d'obtenir une partition du territoire en classes, à l'intérieur desquelles la dispersion des coordonnées géographiques est la plus faible possible tout en ayant un nombre donné d'unités par classe. On présente ici une méthode mise en place initialement pour la constitution d'UP dans le cadre de l'enquête Emploi française (LOONIS 2009), et reprise récemment dans le cadre de travaux pour la constitution d'UP de l'échantillon maître français (ensemble des enquêtes auprès des ménages) (FAVRE-MARTINOZ et al. 2017).

Le principe général est le suivant :

1. On considère le géoréférencement le plus fin possible des unités statistiques. Du fait de la qualité du géoréférencement ou de la nature des données, le nombre n_{xy} d'unités statistiques localisées au point de coordonnées $(x;y)$ peut être strictement supérieur à 1.
2. On fait passer un chemin parmi l'ensemble des localisations connues. On utilise pour cela les méthodes évoquées dans le chapitre 2 : "Codifier la structure de voisinage". Dans la mesure où il n'y a pas de nécessité que le chemin revienne à son point de départ, on privilégie le chemin de Hamilton, qui est le plus court (ce chemin minimise la somme des distances entre deux points consécutifs sans fixer de point de départ ou d'arrivée).
3. Pour construire M zones, on parcourt le chemin depuis son origine en cumulant les quantités n_{xy}. Quand ce total dépasse le seuil $c \simeq \frac{N}{M}$, la première UP est constituée. On reprend alors le processus à partir du premier point non encore visité du chemin.
4. Dans des conditions idéales où c divise N et $n_{xy} = 1$ pour tout couple (x,y), la procédure conduit à des unités primaires homogènes géographiquement et de taille constante. Cette heuristique ne conduit cependant pas à un optimum global. Il convient de prévoir, comme dans toute classification, une procédure de consolidation permettant de gérer les éventuelles situations géographiques atypiques et/ou les tailles d'UP trop éloignées de c. Cette dernière situation peut se présenter, par exemple, quand le dernier point du chemin intégré dans une

UP correspond à une valeur très élevée de n_{xy}, ou pour la dernière UP constituée.

Dans la partie qui suit, on montre une application de cette procédure en se focalisant plus particulièrement sur la question de la construction du chemin quand le nombre d'unités secondaires est important.

10.2.3 Application

La figure 10.1 montre les résultats d'une application de la stratégie générale précédente à la région Alsace (ancienne région, avant la restructuration des régions françaises en 2016). Pour les besoins de l'enquête Emploi en continu (EEC), les résidences principales de la région sont regroupées en unités primaires de 2 600 résidences principales (figure 10.1b), elles-mêmes découpées en secteurs ayant chacun 120 résidences principales (figure 10.1c).

Pour des raisons de temps de calcul dans la constitution des UP, les quelques 616 000 résidences principales ont été initialement regroupées dans 80 000 carreaux de 100 mètres de côté (figure 10.1a), qui constituent donc le géoréférencement le plus fin des unités statistiques. Pour la construction des secteurs au sein des UP, les résidences principales sont par nature géocodées à l'immeuble. La variabilité de la taille des carreaux d'origine ou des immeubles, et donc des n_{xy}, explique en partie la légère variabilité de la taille des UP et de la taille des secteurs finalement obtenues (tableau 10.1).

Ordre du fractile	Taille du carreau d'origine	Taille de l'UP (figure 10.1b)	Taille du secteur (figure 10.1c)
100 %	378	2776	139
99 %	59	2757	131
95 %	23	2685	130
90 %	15	2640	130
75 %	9	2606	124
50 %	5	2595	119
25 %	2	2591	118
10 %	1	2587	118
5 %	1	2502	117
1 %	1	2491	111
0 %	1	2479	99

TABLE 10.1 – Fractiles en nombre de résidences principales dans les carreaux, les UP et les secteurs de la figure 10.1

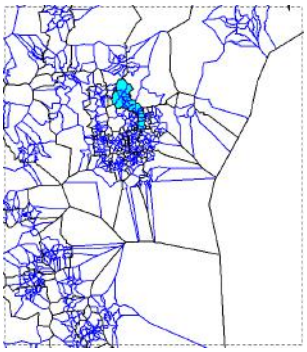
La constitution des secteurs de taille 120 à partir d'un grand nombre de résidences principales peut vite poser des problèmes de temps de calcul. Avec une distance euclidienne, la constitution du chemin de Hamilton le plus court peut être exacte dès lors que le nombre de points n'excède pas quelques centaines. Quand il y en a plusieurs milliers, il n'est pas plus raisonnable qu'utile de chercher à construire le chemin optimal de façon exacte. Nous proposons donc une méthode approchée visant à construire rapidement un chemin qui aura de bonnes propriétés compte tenu de l'objectif fixé. Les différentes étapes de cette méthode approchée sont décrites ci-dessous et illustrées en figure 10.2 pour une UP prise en exemple. Celle-ci comporte 2 600 résidences principales et 1 085 immeubles.

1. Les 1 085 immeubles et 2 600 résidences principales de l'UP en bleu dans la figure 10.1c, sont regroupés par la méthode des **nuées dynamiques** (ou *k-means*) [3] en 20 classes d'effectifs différents mais cohérentes géographiquement. Les variables sur lesquelles s'effectue cette

3. La méthode *k-means* vise à créer des classes homogènes en maximisant la variance entre les classes et en minimisant la variance au sein de chacune d'entre elles.

(a) 616 000 résidences principales, dans 80 000 carreaux de 100 m de côté ...

(b) ... sont réparties dans des UP homogènes de 2 600 résidences principales ...



(c) ... et découpées en secteurs de 120 résidences principales.

FIGURE 10.1 – Construction de zones de faible étendue géographique et de taille constante en nombre de résidences principales en Alsace

classification sont les coordonnées géographiques des immeubles. On notera que $20 \simeq \frac{2600}{120}$ (figures 10.2a et 10.2b).

2. On fait passer un **chemin** de Hamilton parmi les barycentres de ces 20 classes de manière à pouvoir les ordonner (figure 10.2c).
3. Dans une classe donnée i, on **ordonne** les immeubles selon **deux sous-parties** (figure 10.2d) :
   (a) la première regroupe les immeubles de la classe i qui sont plus près de G_{i-1} (barycentre de la classe précédente) que de G_{i+1} (barycentre de la classe suivante), triés par distance croissante à G_{i-1} ;
   (b) la seconde regroupe les immeubles de la classe i qui sont plus près G_{i+1} (barycentre de la classe suivante) que de G_{i-1} (barycentre de la classe précédente), triés par distance décroissante à G_{i+1}.
4. Par construction, les premiers immeubles de la classe i sont proches des derniers immeubles de $i-1$, et les derniers de i sont proches des premiers de $i+1$. Le chemin revient ainsi à parcourir les immeubles par classe, par sous-partie puis par distance croissante ou décroissante selon le cas (figure 10.2e). À l'intérieur d'un immeuble, si besoin, il est possible de trier les logements par étage.

10.3 Comment sélectionner un échantillon spatialement dispersé ?

Les considérations générales ont montré qu'une estimation sera d'autant plus précise que le plan de sondage privilégie la sélection d'individus géographiquement éloignés les uns des autres. GRAFSTRÖM et al. 2013, par exemple, ont formalisé ces considérations de manière plus explicite. Dans cette partie, nous détaillons les méthodes permettant de sélectionner des échantillons spatialement dispersés. Ces méthodes peuvent être regroupées en deux familles.

Pour les méthodes de la première famille, les probabilités d'inclusion des unités sont mises à jour localement afin de limiter la sélection de deux unités voisines. Dans cette famille, on retrouve la méthode du tirage poissonien corrélé spatialement (GRAFSTRÖM 2012), la méthode du pivot spatial (GRAFSTRÖM et al. 2012), et la méthode du cube spatial (GRAFSTRÖM et al. 2013). La deuxième famille se caractérise par une transformation du problème de proximité des unités dans plusieurs dimensions en un problème d'ordre dans $\mathbb{R}$. Ensuite, le tirage s'effectue selon un tirage excluant deux unités proches dans le fichier trié. Cette famille de méthodes regroupe la méthode *general randomized tessellation stratified* (GRTS, STEVENS JR et al. 2004), la méthode basée sur une courbe de Peano (LISTER et al. 2009) ou sur l'algorithme du voyageur de commerce (DICKSON et al. 2016).

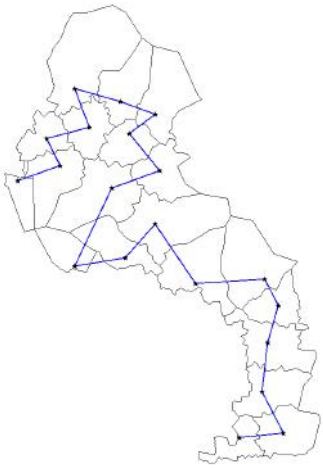
10.3.1 Le tirage poissonien corrélé spatialement (GRAFSTRÖM 2012)

Le tirage poissonien corrélé spatialement est une extension du tirage poissonien corrélé (*Correlated Poisson Sampling*, CPS) proposé par BONDESSON et al. 2008 pour réaliser de l'échantillonnage en temps réel. La méthode CPS est fondée sur un échantillonnage séquentiel et ordonné des unités. Les unités sont ordonnées avec des indices allant de 1 à N. On statue d'abord sur l'unité 1, puis sur l'unité 2, jusqu'à l'unité N. Dans le cas de l'échantillonnage en temps réel, l'ordre de l'indiçage correspond à un ordre de visite préétabli des unités échantillonnables. Dans le cas d'un tirage spatial, l'ordre peut être établi à partir de la proximité des unités selon une fonction de distance euclidienne. À chaque étape, les probabilités d'inclusion sont mises à jour de façon à engendrer une corrélation positive ou négative entre les indicatrices de sélection des unités.

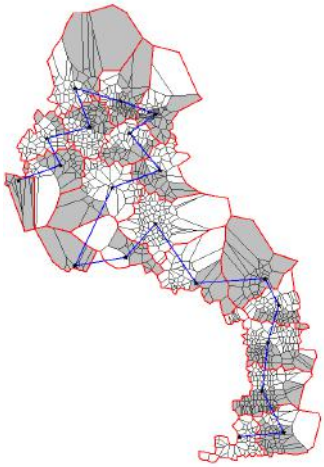
Plus précisément, la première unité est incluse dans l'échantillon avec une probabilité $\pi_1^0 = \pi_1$. On note I_1 l'indicatrice qui vaut 1 si cette unité est sélectionnée, 0 sinon. Plus généralement, à l'étape j, on sélectionne l'unité j avec la probabilité π_j^{j-1} et on met à jour les probabilités

(a) Les immeubles, et leurs cellules de Voronoï associées ...

(b) ... sont regroupés, par nuées dynamiques sur les coordonnées, en une vingtaine de classes de taille variable.



(c) Un chemin passant par les barycentres des classes ...



(d) ... permet de classer les immeubles selon qu'ils sont plus proches du barycentre de la classe précédente (blanc) ou de la classe suivante (gris) ...

(e) ... et donc de créer un chemin passant par tous les immeubles.

(f) En suivant ce chemin, on construit des secteurs de 123 à 128 résidences principales.

FIGURE 10.2 – Des logements aux secteurs de 120 logements

d'inclusion des unités $i \geq j+1$ de la façon suivante :

$$\pi_i^j = \pi_i^{j-1} - (I_j - \pi_j^{j-1}) w_j^i, \tag{10.2}$$

où les w_j^i sont les poids donnés par l'unité j aux unités avec des indices $i \geq j+1$. Les probabilités d'inclusion sont mises à jour étape par étape, avec au plus N étapes jusqu'à l'obtention du vecteur des indicatrices de sélection.

Le choix des poids w_j^i est crucial, car il permet de déterminer si l'on introduit une corrélation positive ou négative entre les indicatrices de sélection. BONDESSON et al. 2008 donnent l'expression de ces poids pour quelques plans de sondage classiques, et une expression générale pour tout plan de sondage. Ainsi, cette méthode est très générale : tout plan de sondage pour lequel les probabilités d'inclusion simple sont fixées peut être implémenté par la méthode CPS. Seules l'expression et les conditions qui portent sur les poids [4] vont varier d'un plan à un autre : par exemple, pour un plan de taille fixe, la somme des poids $w_j^i,\ (j<i)$ doit être égale à 1. Dans un contexte d'autocorrélation spatiale positive (où les unités proches sont semblables), les poids associés doivent être choisis positivement, de façon à introduire une corrélation négative entre les indicatrices de sélection. Il semble donc pertinent de réaliser un test d'autocorrélation spatiale globale pour déterminer le signe des poids à mobiliser dans cette méthode.

GRAFSTRÖM 2012 propose dans son article deux versions pour les poids w_j^i. Nous présentons ici la version considérant une distribution gaussienne. Dans ce cas, les poids sont de la forme :

$$w_j^i \propto \exp(-[d(i,j)/\sigma]^2),\ i = j+1, j+2, ..., N \tag{10.3}$$

.

La constante de proportionnalité est fixée par le fait que la somme des poids doit être égale à 1. Ces poids sont d'autant plus grands que les unités sont proches de l'unité j. Ainsi, la probabilité π_i^j sera d'autant plus faible que l'unité i sera proche (au sens de la distance $d(i,j)$) de l'unité j, ce qui permet de réaliser un échantillonnage spatialement dispersé. Le paramètre σ permet de gérer la dispersion de ces poids, et donc de répartir la mise à jour des probabilités d'inclusion simple dans un voisinage plus ou moins large.

Cette méthode est implémentée sous R dans le package *BalancedSampling* (GRAFSTRÖM et al. 2016) avec la fonction `scps()`.

10.3.2 La méthode du pivot spatial (GRAFSTRÖM et al. 2012)

Rappel sur la méthode du pivot

La méthode du pivot est une procédure d'échantillonnage permettant de sélectionner un échantillon avec des probabilités d'inclusion égales ou inégales (DEVILLE et al. 1998). À chaque étape de l'algorithme, les probabilités d'inclusion de deux unités i et j en lice sont mises à jour, et l'une au moins de ces deux unités est sélectionnée ou rejetée. Le vecteur des probabilités d'inclusion des deux unités en lice (π_i, π_j) est mis à jour selon la règle suivante (combat entre les unités i et j) :

— si $\pi_i + \pi_j < 1$, alors :

$$(\pi_i', \pi_j') = \begin{cases} (0, \pi_i + \pi_j) \text{ avec la probabilité } \dfrac{\pi_j}{\pi_i + \pi_j} \\ (\pi_i + \pi_j, 0) \text{ avec la probabilité } \dfrac{\pi_i}{\pi_i + \pi_j} \end{cases}$$

4. les conditions imposées au poids sont liées aux conditions imposées aux probabilités d'inclusion simple, donc au plan de sondage.

— si $\pi_i + \pi_j \geq 1$, alors :

$$(\pi_i', \pi_j') = \begin{cases} (1, \pi_i + \pi_j - 1) \text{ avec la probabilité } \dfrac{(1-\pi_j)}{(2-\pi_i-\pi_j)} \\ (\pi_i + \pi_j - 1, 1) \text{ avec la probabilité } \dfrac{(1-\pi_i)}{(2-\pi_i-\pi_j)} \end{cases}$$

Cette procédure est répétée jusqu'à l'obtention d'un vecteur des probabilités d'inclusion contenant $N-n$ fois le chiffre 0 et n fois le chiffre 1, déterminant complètement l'échantillon sélectionné (au plus N étapes).

Extension à l'échantillonnage spatial

La méthode du pivot spatial (GRAFSTRÖM et al. 2012) est une extension spatiale de la méthode du pivot. L'idée de la méthode est toujours de mettre à jour le vecteur des probabilités d'inclusion π de façon successive, mais en sélectionnant cette fois-ci à chaque étape deux unités voisines au sens d'une certaine distance (e.g. une distance euclidienne) pour participer au combat. Plusieurs méthodes permettent de sélectionner ces deux unités voisines :

— **LPM1** : deux unités les plus proches l'une de l'autre sont sélectionnées pour participer au combat, *i.e.* une unité i est sélectionnée aléatoirement parmi les N unités de la population, puis l'unité j la plus proche de i est sélectionnée pour participer au combat si et seulement si i est aussi l'unité la plus proche de j (au mieux N^2 étapes, au pire N^3 étapes) ;
— **LPM2** : deux unités voisines sont sélectionnées pour participer au combat, *i.e.* une unité i est sélectionnée aléatoirement parmi les N unités de la population, puis l'unité j la plus proche de i est sélectionnée pour participer au combat (N^2 étapes) ;
— **LPM K-D TREE** : les deux unités voisines sont sélectionnées au moyen d'un arbre *k-d* de partition de l'espace (LISIC 2015) permettant de faire les recherches de plus proches voisins plus rapidement (complexité de l'algorithme en $N\log(N)$).

Ces trois méthodes du pivot spatial sont implémentées en C++ dans le package *BalancedSampling* du logiciel R.

10.3.3 La méthode du cube

Généralités sur la méthode du cube

Le tirage équilibré est une procédure dont le but est de fournir un échantillon respectant les deux contraintes suivantes :

— les probabilités d'inclusion sont respectées ;
— l'échantillon est équilibré sur p variables auxiliaires. Autrement dit les estimateurs de Narain-Horvitz-Thompson des totaux des variables auxiliaires sont égaux aux totaux de ces variables auxiliaires dans la population :

$$\sum_{i \in \mathbb{S}} \frac{x_i}{\pi_i} = \sum_{i \in U} x_i \tag{10.4}$$

Un algorithme permettant de réaliser un tel tirage est l'algorithme du cube. Pour en décrire le principe, il est opportun d'avoir recours à la représentation géométrique suivante. Un échantillon est un des sommets d'un hypercube de dimension N, noté C. L'ensemble des p contraintes, rappelées par l'équation (10.4), définit un hyperplan de dimension $N-p$, noté Q. On note $K = Q \cap C$, l'intersection du cube et de l'hyperplan. Une représentation graphique du problème en dimension 3, tirée de l'article DEVILLE et al. 2004, est donnée ci-dessous (figure 10.3).

L'algorithme du cube se décompose en deux phases. La première phase, dite "phase de vol" (figure 10.4), est une marche aléatoire qui part du vecteur des probabilités d'inclusion et évolue dans K. Pour cela, on part de $\pi(0) = \pi$, puis on met à jour le vecteur des probabilités d'inclusion

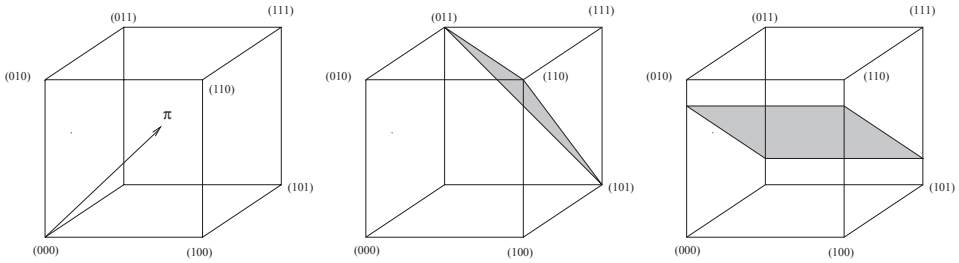


FIGURE 10.3 – Représentation graphique du cube pour $N = 3$ et différentes configurations possibles de l'espace des contraintes avec ici $p = 1$

en choisissant un vecteur $u(0)$ de sorte que $\pi + u(0)$ demeure dans l'espace des contraintes. En suivant la direction donnée par le vecteur $u(0)$, on aboutit nécessairement sur une face du cube. Le sens pour la mise à jour du vecteur des probabilités d'inclusion est ensuite donné par les paramètres $\lambda_1^*(0)$ et $\lambda_2^*(0)$, ceux-ci étant choisis de sorte que le vecteur mis à jour $\pi(1)$ touche une face du cube. Le choix du sens pour cette mise à jour est effectué de façon aléatoire de façon que $E(\pi(1)) = \pi(0)$. On recommence ensuite l'opération en choisissant un nouveau vecteur $u(1)$ pour la direction et un nouveau sens pour la mise à jour des probabilités d'inclusion. Cette marche aléatoire s'arrête lorsqu'elle a atteint un sommet π^* de K. À l'issue de cette première phase, le sommet π^* n'est pas nécessairement un sommet du cube C. Soit q, le nombre de composantes non entières dans le vecteur π^* ($q \leq p$). Si q est nul, la procédure d'échantillonnage est terminée, sinon il faut procéder à la deuxième étape, appelée "phase d'atterrissage". Elle consiste à relâcher le moins possible les contraintes d'équilibrage, et à relancer une phase de vol avec ces nouvelles contraintes jusqu'à l'obtention d'un échantillon. Il n'est pas envisageable de modifier dès le début l'espace des contraintes de sorte que les sommets de K soient confondus avec ceux de C, car cela reviendrait à tester tous les échantillons possibles pour voir dans un premier temps si l'un d'eux permet de respecter les contraintes. Le fait de modifier l'espace des contraintes dans un deuxième temps (dans la phase d'atterrissage) permet de travailler sur une population U^* de plus petite taille ($\dim(U^*) = q$). Le problème peut ainsi être résolu car le nombre d'échantillons à considérer est raisonnable.

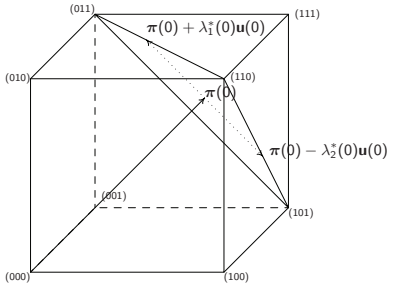


FIGURE 10.4 – Première étape de la phase de vol du cube pour $N = 3$ et une contrainte ($p = 1$) d'un échantillon de taille fixe $n = 2$

L'implémentation de cet algorithme est disponible sous SAS grâce à la macro *FAST CUBE* ou sous R dans le package *BalancedSampling*.

La méthode du cube spatial

L'idée générale de l'algorithme de tirage spatialement équilibré est de construire un cluster de $p+1$ unités géographiquement proches, puis d'appliquer la phase de vol du cube sur ce cluster. Cela

conduit à statuer sur la sélection ou non d'une unité dans ce cluster en respectant les p contraintes localement dans ce cluster. Ensuite, les probabilités sont modifiées localement, ce qui assure que les probabilités d'inclusion des unités proches sont réduites si l'unité sur laquelle on a statué est sélectionnée. Cela limite ainsi la probabilité qu'une de ses unités proches soit sélectionnée dans l'étape suivante de l'algorithme. Puis on répète la procédure : on sélectionne une unité, on crée un cluster de $p+1$ unités autour de l'unité sélectionnée, et on applique la phase de vol du cube avec les probabilités d'inclusion mises à jour à l'étape précédente. On répète le processus tant qu'il reste plus de $p+1$ unités. Pour finir, on applique la phase d'atterrissage classique du cube.

La méthode de tirage spatialement équilibrée décrite ci-dessus est disponible dans le package R *BalancedSampling*. Ce package, développé en C++, permet d'appliquer l'algorithme très rapidement.

Équilibrage sur les moments

La définition d'un échantillon spatialement équilibré laisse entrevoir une utilisation différente de l'algorithme du cube pour l'échantillonnage spatial. Pour MARKER et al. 2009, "un échantillon est spatialement équilibré si les moments spatiaux des échantillons localisés correspondent aux moments spatiaux de la population. Les moments spatiaux sont le centre de gravité et l'inertie.". Dans la terminologie de l'algorithme du cube, cette définition peut revenir à sélectionner un échantillon équilibré sur des variables définies à partir des coordonnées géographiques : $x_i, y_i, x_i^2, y_i^2, x_i y_i$ afin de respecter les moments non centrés d'ordre 1 et 2 :

— $T_x = \sum_{i \in U} x_i$,
— $T_y = \sum_{i \in U} y_i$,
— $T_{x^2} = \sum_{i \in U} x_i^2$,
— $T_{y^2} = \sum_{i \in U} y_i^2$,
— $T_{xy} = \sum_{i \in U} x_i y_i$.

10.3.4 Méthodes sur fichier trié

Les méthodes existantes dans cette famille (STEVENS JR et al. 2004, DICKSON et al. 2016, LISTER et al. 2009) s'appuient dans un premier temps sur la constitution d'un chemin passant par toutes les unités statistiques. Ce chemin peut être GRTS (*generalized random tessellation stratified*), voyageur de commerce (TSP), ou une courbe de Peano. Conditionnellement à l'ordre défini par ce chemin, il s'agit ensuite de sélectionner un échantillon selon une méthode qui exclut deux unités proches, par exemple le tirage systématique.

D'autres méthodes de constitution de chemin existent (chemins de Hamilton, ou courbes remplissant l'espace : Hilbert, Lebesgue). De même, il existe d'autres méthodes de sélection excluant les unités proches à tri donné, comme celle des plans de sondage déterminantaux (LOONIS et al. 2018). La question des chemins ayant été abordée dans le chapitre 2 : "Codifier la structure de voisinage", on présente ici les propriétés répulsives des plans systématiques et déterminantaux.

La méthode de tirage systématique

Le tirage systématique est une méthode de tirage simple à mettre en œuvre et qui permet de réaliser un tirage à probabilités inégales tout en respectant les probabilités d'inclusion simple. Cette méthode a été proposée par MADOW 1949, puis étendue par CONNOR 1966, BREWER 1963, PINCIARO 1978, et HIDIROGLOU et al. 1980. Il est très souvent utilisé en pratique pour les enquêtes par téléphone, pour faire de l'échantillonnage sur des flux continus de données, ou dans le tirage des logements dans le cas des enquêtes ménages de l'Insee.

Pour tirer un échantillon de taille fixe n respectant le vecteur de probabilités d'inclusion π, on commence par définir la somme cumulée des probabilités d'inclusion par $V_i = \sum_{l=1}^{i} \pi_l, i \in U$, avec $V_0 = 0$. Pour un échantillon de taille fixe, on a $V_N = n$. Ensuite, on utilise l'algorithme du tirage systématique présenté ci-dessous pour statuer sur les unités à échantillonner.

Algorithme du tirage systématique :

— Générer une variable aléatoire u uniformément distribuée sur l'intervalle $[0,1]$.

— Pour $i = 1,...,N$,

$$I_i = \begin{cases} 1 & \text{si il existe un entier } j \text{ tel que } V_{i-1} \leq u + j - 1 < V_i, \\ 0 & \text{sinon.} \end{cases}$$

Le tableau 10.2 présente un exemple de la méthode dans le cas où $n = 3$ et $N = 10$.

i	1	2	3	4	5	6	7	8	9	10
π_i	0.2	0.2	0.3	0.3	0.4	0.4	0.3	0.3	0.3	0.3
V_i	0.2	0.4	0.7	1	1.4	1.8	2.1	2.4	2.7	3(=n)

TABLE 10.2 – Un exemple de tirage systématique

Par exemple si le nombre aléatoire généré u est égal à 0.53, les unités 2, 5, 8 seront sélectionnées car satisfaisant les contraintes :

$$V_2 \leq u < V_3 \qquad V_5 \leq u+1 < V_6 \qquad V_8 \leq u+2 < V_9.$$

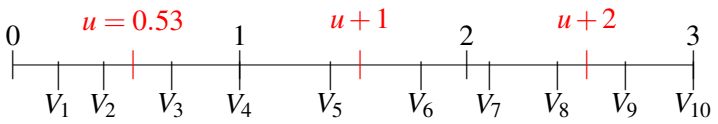


FIGURE 10.5 – Sélection de 3 unités parmi 10

Selon cette méthode, les unités (i, j) telles que $|V_i - V_j| < 1$ ont une probabilité nulle d'être sélectionnées ensemble. Si le fichier est trié judicieusement, cela assure la dispersion géographique de l'échantillon.

Implémentation de la méthode GRTS

Parmi les pratiques de tirage systématique sur fichier trié géographiquement, la méthode GRTS est très utilisée. Le tri GRTS est décrit dans le chapitre 2 : "Codifier la structure géographique". Le package *gstat* du logiciel R a été implémenté spécifiquement pour tirer des échantillons à l'aide de cette méthode. Cependant, la méthode GRTS présente quelques désavantages, en particulier le fait que l'algorithme de découpage et celui de tirage ne soient pas dissociés, ou encore le temps d'exécution de la méthode. En effet, la méthode propose par défaut de s'arrêter à 11 niveaux hiérarchiques pour le découpage, le temps d'exécution de la méthode risquant d'être trop important si un découpage plus fin était demandé. Au-delà, si plus d'une unité appartiennent à la même cellule, un échantillon est sélectionné au sein de la cellule à l'aide de la fontion *S-PLUS*, prenant pour paramètre *prob* qui correspond aux probabilités d'inclusion de chaque élément. L'algorithme de GRTS s'adapte ainsi difficilement à des populations de grande taille. Afin de pallier ces limites d'ordre computationnel, une nouvelle méthode du Pivot par Tessellation (CHAUVET et al. 2017) a

été développée en R, où l'algorithme de Tessellation (très proche de celui de GRTS) est dissocié de la partie tirage. Cette méthode s'appuie sur la décomposition binaire d'un nombre, ce qui permet d'effectuer le découpage directement sur 31 niveaux. Le temps d'exécution est donc considérablement amélioré. De plus, il est possible d'effectuer ce découpage dans plus de deux dimensions.

Échantillonnage déterminantal

Par définition, une variable aléatoire $\mathbb{S}$ à valeurs dans 2^U a pour loi de probabilité un plan de sondage déterminantal si et seulement si il existe une matrice hermitienne contractante [5] K indexée par U, appelée noyau, telle que pour tout $s \in 2^U$,

$$p(s \subseteq \mathbb{S}) = \det(K_{|s}) \tag{10.5}$$

où $K_{|s}$ est la sous matrice de K indexée par les unités de s. De cette définition découle directement le calcul des probabilités d'inclusion (table 10.3).

π_i	=	$pr(i \in \mathbb{S})$	=	$\det(K_{\|\{i\}})$	=	K_{ii}
π_{ij}	=	$pr(i,j \in \mathbb{S})$	=	$\det\begin{pmatrix} K_{ii} & K_{ij} \\ \overline{K}_{ij} & K_{jj} \end{pmatrix}$	=	$K_{ii}K_{jj} - \mid K_{ij} \mid^2$

TABLE 10.3 – Calcul des probabilités d'inclusion simple et double dans un plan de sondage déterminantal de noyau K
Note : $|z|$ désigne le module du nombre complexe z.

La diagonale de la matrice K correspond aux probabilités d'inclusion simple. Un autre résultat particulièrement important des plans déterminantaux est celui précisant qu'un plan déterminantal est de taille fixe si et seulement si K est une matrice de projection [6](HOUGH et al. 2006).

On considère l'ensemble des matrices de projection dont la diagonale correspond à un vecteur Π de probabilités d'inclusion fixées *a priori*. Parmi elles, la matrice K^Π (dont les coefficients sont donnés par la table 10.4) présente des propriétés intéressantes en termes de répulsion spatiale.

	Valeurs de j	
Valeurs de i	$j = i_r$	$i_r < j < i_{r+1}$
$i_{r'} < i < i_{r'+1}$	$-\sqrt{\Pi_i}\sqrt{\frac{(1-\Pi_j)(\Pi_j-\alpha_l)}{1-(\Pi_j-\alpha_j)}}\gamma_r^{r'}$	$\sqrt{\Pi_i\Pi_j}\gamma_r^{r'}$
$i = i_{r'+1}$	$-\sqrt{\frac{(1-\Pi_i)\alpha_i}{1-\alpha_i}}\sqrt{\frac{(1-\Pi_j)(\Pi_j-\alpha_j)}{1-(\Pi_j-\alpha_j)}}\gamma_r^{r'}$	$\sqrt{\frac{(1-\Pi_i)\alpha_i}{1-\alpha_i}}\sqrt{\Pi_j}\gamma_r^{r'}$

où pour tout r tel que $1 \leq r \leq n$:

— $1 < i_r \leq N$ est un entier tel que $\sum_{i=1}^{i_r-1} \Pi_i < r$ et $\sum_{i=1}^{i_r} \Pi_i \geq r$; par convention on posera $i_0 = 0$

— $\alpha_{i_r} = r - \sum_{i=1}^{i_r-1} \Pi_i$. On notera que $\alpha_{i_r} = \Pi_{i_r}$ si $\sum_{i=1}^{i_r} \Pi_i = r$.

— $\gamma_r^{r'} = \sqrt{\prod_{k=r+1}^{r'} \frac{(\Pi_{i_k}-\alpha_{i_k})\alpha_{i_k}}{(1-\alpha_{i_k})(1-(\Pi_{i_k}-\alpha_{i_k}))}}$ pour $r < r'$, $\gamma_r^{r'} = 1$ autrement.

TABLE 10.4 – Valeurs de K_{ij}^Π avec $i > j$

5. Une matrice complexe K est hermitienne si $K = \overline{K}^t$, où les coefficients de $\overline{K}$ sont les conjugués de ceux de K. Une matrice est contractante si toutes ses valeurs propres sont comprises entre 0 et 1.

6. Une matrice hermitienne est de projection si ses valeurs propres sont 0 ou 1.

La répulsion du plan déterminantal associé à K^{Π} pour des individus proches (selon l'ordre du fichier) est illustrée par les propriétés suivantes (LOONIS et al. 2018) :

1. le plan sélectionnera au plus un individu dans un intervalle de la forme $]i_r+1, i_{r+1}-1[$;
2. si un individu y est tiré, ainsi que l'individu "proche" i_{r+1}, alors le plan ne sélectionnera pas d'individu supplémentaire "proche", c'est-à-dire dans $]i_{r+1}+1, i_{r+2}-1[$;
3. ce plan aura toujours au moins un individu dans un intervalle $[i_r+1, i_{r+1}-1]$;
4. si $|i-j|$ est grand, alors $\pi_{ij} \approx \Pi_i \Pi_j$. On retrouve les probabilités d'inclusion double du plan poissonien.

L'application des résultats sur les plans déterminantaux aux probabilités définies dans la table 10.2 conduit aux quantités : $i_1 = 4, i_2 = 7, i_3 = 10$ et $\alpha_4 = 0.3 = \Pi_4, \alpha_7 = 0.2, \alpha_{10} = 0.3 = \Pi_{10}$. Les probabilités d'inclusion doubles sont données par la matrice ci-dessous.

$$\begin{pmatrix}
 & 0 & 0 & 0 & \frac{2}{25} & \frac{2}{25} & \frac{3}{50} & \frac{3}{50} & \frac{3}{50} & \frac{3}{50} \\
0 & & 0 & 0 & \frac{2}{25} & \frac{2}{25} & \frac{3}{50} & \frac{3}{50} & \frac{3}{50} & \frac{3}{50} \\
0 & 0 & & 0 & \frac{3}{25} & \frac{3}{25} & \frac{9}{100} & \frac{9}{100} & \frac{9}{100} & \frac{9}{100} \\
0 & 0 & 0 & & \frac{3}{25} & \frac{3}{25} & \frac{9}{100} & \frac{9}{100} & \frac{9}{100} & \frac{9}{100} \\
\frac{2}{25} & \frac{2}{25} & \frac{3}{25} & \frac{3}{25} & & 0 & \frac{1}{20} & \frac{7}{60} & \frac{7}{60} & \frac{7}{60} \\
\frac{2}{25} & \frac{2}{25} & \frac{3}{25} & \frac{3}{25} & 0 & & \frac{1}{20} & \frac{7}{60} & \frac{7}{60} & \frac{7}{60} \\
\frac{3}{50} & \frac{3}{50} & \frac{9}{100} & \frac{9}{100} & \frac{1}{20} & \frac{1}{20} & & \frac{1}{15} & \frac{1}{15} & \frac{1}{15} \\
\frac{3}{50} & \frac{3}{50} & \frac{9}{100} & \frac{9}{100} & \frac{7}{60} & \frac{7}{60} & \frac{1}{15} & & 0 & 0 \\
\frac{3}{50} & \frac{3}{50} & \frac{9}{100} & \frac{9}{100} & \frac{7}{60} & \frac{7}{60} & \frac{1}{15} & 0 & & 0 \\
\frac{3}{50} & \frac{3}{50} & \frac{9}{100} & \frac{9}{100} & \frac{7}{60} & \frac{7}{60} & \frac{1}{15} & 0 & 0 &
\end{pmatrix}.$$

Les termes autour de la diagonale principale ont tendance à être nuls ou proches de 0, traduisant la répulsion.

10.4 Comparaison des méthodes

Différentes méthodes d'échantillonnage visant à prendre en compte l'information spatiale ont été présentées. Cette partie s'attache à comparer leur efficacité relative, en s'appuyant sur des données réelles.

10.4.1 Le principe

Les données fiscales exhaustives de 2015 sont géoréférencées pour l'ensemble des ménages, ce qui autorise le partitionnement du territoire de la région Provence-Alpes-Côte d'Azur (PACA) en 1 012 unités primaires (UP) d'environ 2 000 résidences principales. Chacune de ces UP est caractérisée par une quinzaine de variables d'intérêt décrivant sa situation socio-économique ou démographique. On s'intéresse aux propriétés statistiques de l'échantillonnage au premier degré de tirage, c'est-à-dire un tirage de m unités primaires parmi les $M = 1\,012$ UP.

On se place dans la situation où l'on ne dispose que des coordonnées géographiques du barycentre des UP au moment de la sélection de l'échantillon. On teste deux jeux de probabilités d'inclusion : le premier correspond à des probabilités constantes, le second à des probabilités proportionnelles au nombre de chômeurs. On teste ces deux jeux pour trois tailles différentes d'échantillons : $m = 30, 60, 100$.

On cherche à évaluer les méthodes présentées précédemment en comparant leurs performances à celles d'une méthode *benchmark* : il s'agit du sondage aléatoire simple pour les plans à probabilités constantes et du tirage systématique trié aléatoirement pour ceux à probabilités inégales. La performance d'une méthode est mesurée à l'aune de deux types d'indicateurs :

1. **variances d'estimateurs de totaux**
   Pour chaque méthode, on cherche à voir dans quelle mesure la variance du total d'une variable

donnée est diminuée par rapport à la variance obtenue avec la méthode *benchmark*. On étudie cela pour un ensemble de variables d'intérêt présentant différents niveaux d'autocorrélation spatiale.
Pour toutes les méthodes sauf celle des plans déterminantaux, les variances des totaux sont estimées par méthode de Monte Carlo, en répliquant 10 000 fois chaque méthode pour chaque jeu de probabilités d'inclusion et chaque taille d'échantillon. Pour les plans déterminantaux, les probabilités d'inclusion double étant connues, la variance peut être calculée de manière exacte.
Puisque l'on souhaite savoir si le gain en variance est plus important quand la variable est autocorrélée spatialement, les 15 variables d'intérêt sont hiérarchisées en fonction de leur niveau d'autocorrélation spatiale, mesuré par l'indice de Moran dilaté par les probabilités d'inclusion. C'est en effet la quantité $\frac{y_i}{\pi_i}$ qui conditionne la qualité des résultats. Lorsque le plan est à probabilités constantes, cela revient à calculer l'indice de Moran directement pour chaque variable (table 10.5).

2. **indicateur dit de Voronoï**
Pour chaque méthode, on calcule également un indicateur empirique de dispersion (dit indice de Voronoï), en suivant STEVENS JR et al. 2004. Son principe est le suivant :
   — on construit le diagramme de Voronoï associé aux seules m UP sélectionnées ;
   — pour une UP i échantillonnée, on identifie, parmi les 1 012 UP d'origine, celles situées dans la cellule associée à i ;
   — on calcule la somme δ_i des probabilités d'inclusion de ces UP. La moyenne des δ_i est égale à 1, puisque la somme des probabilités d'inclusion sur les 1 012 UP vaut m et que les m cellules partitionnent l'espace. Si la procédure a sélectionné peu d'unités autour d'une UP i donnée, δ_i sera supérieur à 1. Si la procédure a sélectionné d'autres unités proches de i, δ_i sera inférieur à 1 (voir figure 10.6) ;
   — pour un échantillon aléatoire $\mathbb{S}$, on définit alors l'indicateur de Voronoï par :

$$\Delta_{\mathbb{S}} = \frac{1}{m-1} \sum_{i \in \mathbb{S}} (\delta_i - 1)^2.$$

Plus une procédure répartit uniformément au plan spatial, plus la dispersion des δ_i mesurée par $\Delta_{\mathbb{S}}$ sera faible. L'espérance de $\Delta_{\mathbb{S}}$ sera estimée par simulation (moyenne sur 10 000 réplications, notée V).

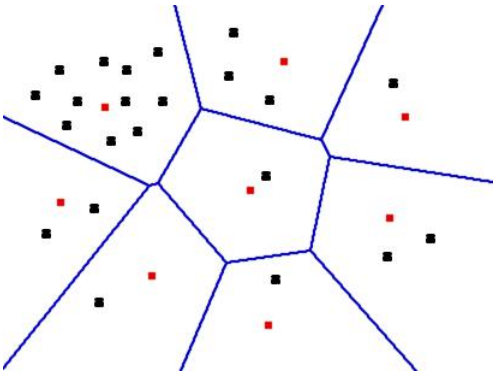


FIGURE 10.6 – Calcul de l'indice de Voronoï
Note : les cellules sont construites autour des unités sélectionnées (rouge). Les δ_i sont calculés sur l'ensemble des unités (rouge et noir)

L'indicateur de Voronoï peut être calculé en R en utilisant la fonction `sb()` du package

BalancedSampling ou à partir des codes R donnés dans BENEDETTI et al. 2015 (pp. 161-162).

Variable	**I de Moran** π constant	**I de Moran dilaté** ($\frac{y_i}{\pi_i}$)
Nombre de ménages percevant des revenus agricoles	0,68	0,66
Revenus salariaux totaux	0,62	0,55
Nombre de couples avec enfant(s)	0,61	0,54
Nombre de bénéficiaires de minima sociaux	0,60	0,61
Nombre de pauvres	0,58	0,58
Nombre d'enfants	0,55	0,52
Nombre d'individus vivant dans un quartier politique de la ville	0,55	0,54
Nombre de ménages propriétaires	0,52	0,47
Niveau de vie total	0,46	0,46
Nombre de chômeurs	0,45	0,42
Nombre de famille monoparentales	0,41	0,43
Nombre d'individus	0,40	0,34
Nombre d'hommes	0,39	0,34
Nombre de femmes	0,24	0,34
Nombre de ménages	0,08	0,38

TABLE 10.5 – Indices de Moran pour différentes variables calculées au niveau des UP de la région PACA
Source : *Insee, Fideli 2015*

10.4.2 Résultats

Une dizaine de méthodes d'échantillonnage spatial sont étudiées :

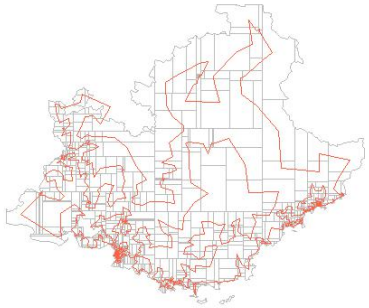
— 4 méthodes de la famille des méthodes de mises à jour itératives des probabilités d'inclusion, dite famille A dans la suite : tirage poissonien, pivot spatial, cube spatial [7], et cube équilibré sur les moments spatiaux ;

— 6 méthodes de la seconde famille, dite famille B dans la suite, fondée sur des tris préalables du fichier. En effet, 3 chemins sont envisagés (figure 10.7) : le chemin du voyageur de commerce (10.7a), celui de Hamilton (10.7b), et GRTS (10.7c), et chacun est suivi d'un tirage systématique ou de la méthode des plans déterminantaux. Les trois chemins sont obtenus à partir d'une méthode exacte et les réplications de tirages d'échantillons ont donc lieu sur un fichier trié de façon unique.

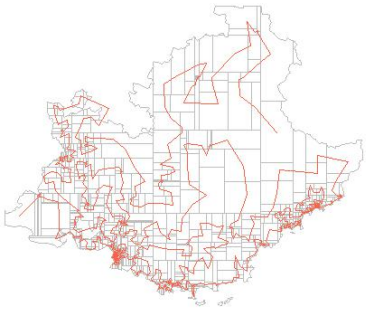
La figure 10.8 représente la quantité $(V^q - V^{ref})/V^{ref}$, où V^q est l'indicateur de Voronoï pour la méthode q et V^{ref} le même indicateur pour le *benchmark*. Une valeur fortement négative révèle une meilleure dispersion spatiale. La figure montre que, pour toutes les méthodes et toutes les tailles d'échantillon, l'indicateur de Voronoï est sensiblement amélioré : de -60 à -70 % par rapport au *benchmark*. La méthode des moments équilibrés est moins performante, tout en restant meilleure que le *benchmark* cependant.

Pour une méthode et une taille d'échantillon données, les figures 10.10 et 10.9 représentent, comme pour l'indicateur de Voronoï, la diminution, par rapport au *benchmark*, de la variance d'une

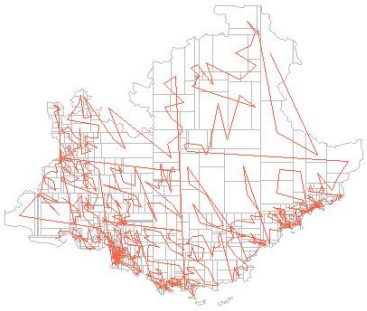
7. Le pivot spatial et le cube spatial sont deux méthodes équivalentes dans ce contexte.



(a) Voyageur de commerce (TSP)



(b) Hamilton



(c) GRTS

FIGURE 10.7 – Chemins reliant les centroïdes des UP
Source : *Insee, Fideli 2015*

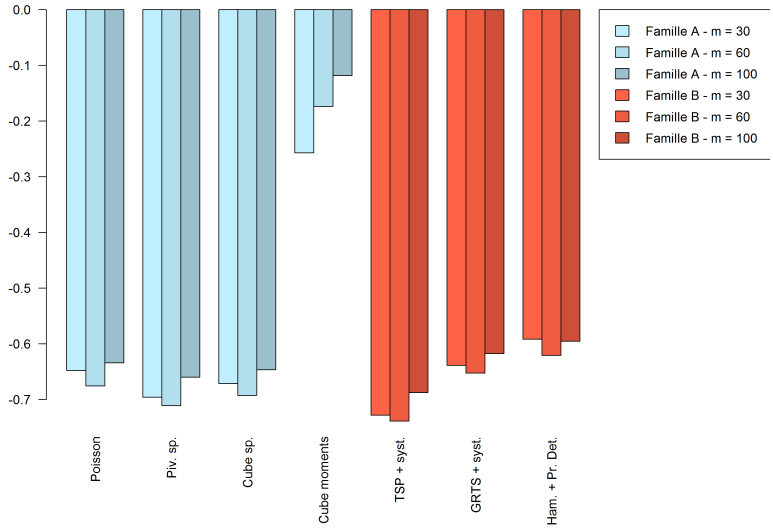


FIGURE 10.8 – $(V^q - V^{ref})/V^{ref}$, où V^q est l'indicateur de Voronoï pour la méthode q et V^{ref} pour le *benchmark*, pour différentes valeurs de m (exemple des probabilités égales)
Source : *Insee, Fideli 2015*
Note : Pour un tirage de 30 UP selon la méthode du tirage poissonien à probabilités égales, l'indicateur de Voronoï (moyenné sur 10 000 réplications) est diminué de 65 % par rapport au sondage aléatoire simple (*benchmark*).

variable d'intérêt. Cette diminution est mise en relation avec l'intensité de l'autocorrélation spatiale de la variable déflatée des probabilités d'inclusion.

Pour les méthodes représentées en figure 10.10, soit la plupart des méthodes étudiées, le gain en termes de variance est d'autant plus fort que la variable est autocorrélée spatialement. Ce résultat est néanmoins plus net à probabilités constantes (10.10a) qu'à probabilités inégales (10.10b). Ces méthodes sont équivalentes en termes de gain. Ainsi le tirage poissonien, le pivot spatial, le cube spatial, et les plans déterminantaux sur fichier trié (tri TSP ou Hamilton), réduisent tous la variance de l'échantillon presque de moitié pour les variables les plus autocorrélées et pour $m = 100$. Par ailleurs, pour toutes les méthodes représentées dans la figure 10.10, le gain relatif en variance est d'autant plus fort que le taux de sondage est élevé. La figure 10.11 illustre ce résultat pour les plans déterminantaux à probabilités constantes.

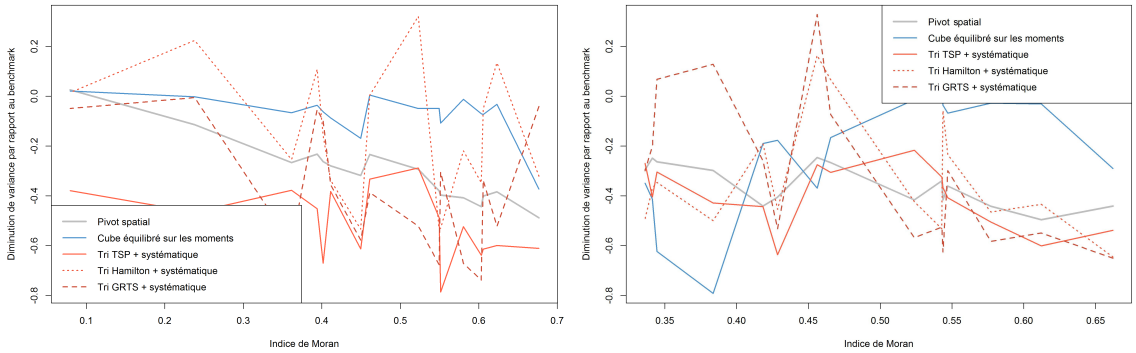
Les quatre méthodes représentées en rouge et bleu dans la figure 10.9 se distinguent par leurs résultats :

— la méthode du cube équilibré sur les moments spatiaux d'ordre 1 et 2 (x, y, $x*y$, x^2 et y^2, où x et y sont les coordonnées spatiales) est, en général, moins performante en termes de gain de variance. En cherchant à se calibrer sur l'inertie de la population totale, cette méthode revient finalement à reproduire dans l'échantillon les regroupements et les éloignements d'unités, de façon antagoniste à la volonté de dispersion de l'échantillon ;
— les tris de fichiers (TSP, Hamilton ou GRTS) suivis d'un tirage systématique, donnent des résultats plus erratiques que les autres méthodes de la même famille. L'entropie [8] du plan de sondage systématique est très faible, et l'est d'autant plus sur un fichier trié de façon unique. Le nombre d'échantillons possibles avec cette méthode est M/m, ce qui explique que ces courbes aient une allure moins lisse et que les conclusions soient plus difficiles à tirer. Ces méthodes restent néanmoins très performantes en termes de dispersion d'échantillon. En particulier, le tri TSP suivi d'un tirage systématique est la méthode qui réduit le plus l'indicateur de Voronoï (figure 10.8). C'est aussi celle qui diminue le plus la variance des variables les plus autocorrélées spatialement. Le tri GRTS, lui, est moins performant, en lien avec une moindre qualité du tri (la longueur totale du chemin obtenu avec GRTS est quasiment deux fois plus grande que le chemin TSP ou Hamilton, voir figure (10.7)).

Conclusion

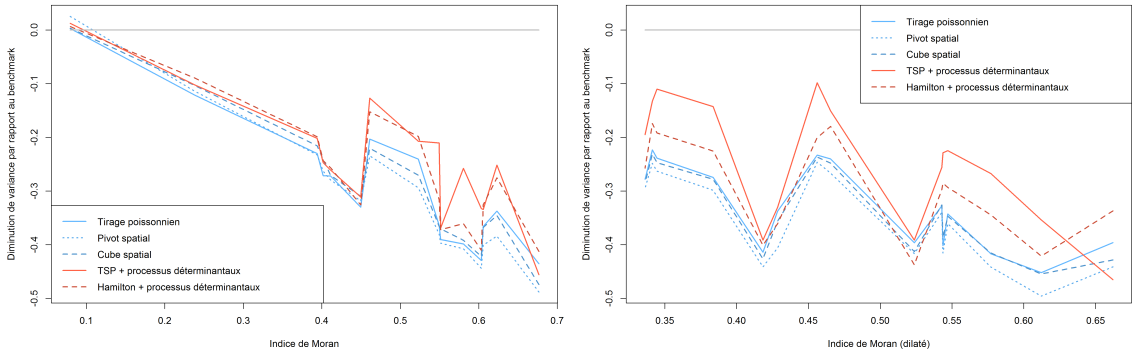
La constitution d'échantillons à partir d'une base de sondage géoréférencée est un nouveau contexte possible de mobilisation judicieuse de l'information géographique. Ce chapitre a présenté différentes méthodes qui utilisent cette information à différents stades de la conception du plan de sondage. En s'appuyant sur des données réelles, nous avons comparé ces différentes méthodes à l'aune d'indicateurs de précision classiques ou originaux, en testant différents jeux de paramètres. La grande majorité des méthodes suggérées s'avèrent efficaces en termes de précision des estimations, même si les méthodes de tirage systématique sur fichier trié apparaissent moins performantes. L'efficacité statistique d'une méthode d'échantillonnage spatial augmente avec le niveau d'autocorrélation spatiale de la variable d'intérêt à estimer.

8. L'entropie est une mesure de désordre. Un plan à forte entropie autorise la sélection d'un grand nombre d'échantillons et laisse donc une place importante à l'aléatoire



(a) Probabilités constantes (b) Probabilités inégales

FIGURE 10.9 – Réductions de variances par rapport au *benchmark* pour différentes méthodes, selon l'indice d'autocorrélation spatiale de la variable (exemple avec $m = 60$)
Note : La méthode du pivot spatial de la figure 10.10 est représentée en trait gris pour comparaison.



(a) Probabilités constantes (b) Probabilités inégales

FIGURE 10.10 – Réductions de variances par rapport au *benchmark* pour différentes méthodes, selon l'indice d'autocorrélation spatiale de la variable (exemple avec $m = 60$)
Source : *Insee, Fideli 2015*
Note : Chaque courbe correspond à une méthode d'échantillonnage spatial, et chaque point de la courbe correspond à 10 000 échantillons tirés selon une même méthode. On représente la variation de la variance d'une variable par rapport à un *benchmark* (en pourcentage), en fonction du niveau d'autocorrélation spatiale de cette variable. Par exemple, pour un tirage à probabilités égales de 60 UP avec la méthode du tirage poissonnien, la variance de la variable "nombre de femmes par UP" (d'indice de Moran 0.24) est diminuée de 11 % par rapport au tirage aléatoire simple

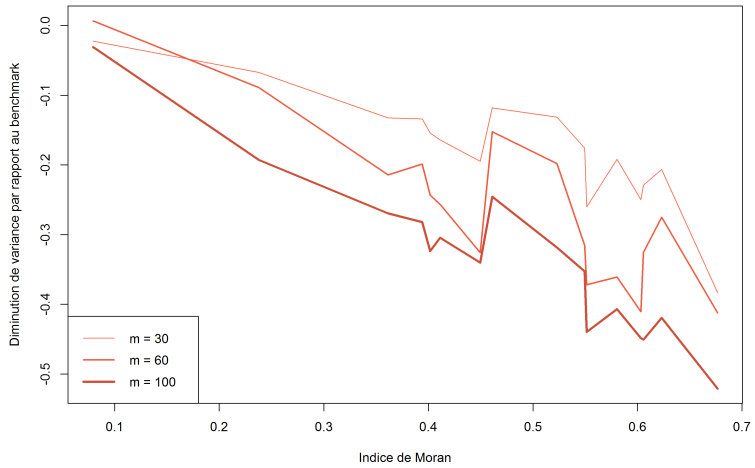


FIGURE 10.11 – Réductions de variances par rapport au *benchmark* selon l'indice d'autocorrélation spatiale de la variable, pour différentes valeurs de m (exemple des processus déterminantaux sur tri de Hamilton, à probabilités constantes)
Source : *Insee, Fideli 2015*
Note : Pour la méthode des processus déterminantaux à probabilités égales, la variance de la variable "nombre de femmes par UP" (d'indice de Moran 0.24) est diminuée de 6,7 % pour un échantillon de 30 UP, de 8,9 % pour un échantillon de 60 UP, et de 19,3 % pour un échantillon de 100 UP, par rapport au tirage aléatoire simple

Références - Chapitre 10

ARDILLY, Pascal (2006). *Les techniques de sondage*. Editions Technip.

BENEDETTI, Roberto, Federica PIERSIMONI et Paolo POSTIGLIONE (2015). *Sampling Spatial Units for Agricultural Surveys*. Springer.

BONDESSON, Lennart et Daniel THORBURN (2008). « A List Sequential Sampling Method Suitable for Real-Time Sampling ». *Scandinavian Journal of Statistics* 35.3, p. 466–483.

BREWER, K.R.W. (1963). « A model of systematic sampling with unequal probabilities ». *Australian & New Zealand Journal of Statistics* 5.1, p. 5–13.

CHAUVET, Guillaume et Ronan LE GLEUT (2017). « Asymptotic results for pivotal sampling with application to spatial sampling ». *Work in progress*.

CONNOR, W.S. (1966). « An exact formula for the probability that two specified sampling units will occur in a sample drawn with unequal probabilities and without replacement ». *Journal of the American Statistical Association* 61.314, p. 384–390.

DEVILLE, Jean-Claude et Yves TILLÉ (1998). « Unequal probability sampling without replacement through a splitting method ». *Biometrika* 85.1, p. 89–101.

— (2004). « Efficient balanced sampling : the cube method ». *Biometrika* 91.4, p. 893–912.

DICKSON, Maria Michela et Yves TILLÉ (2016). « Ordered spatial sampling by means of the traveling salesman problem ». *Computational Statistics*, p. 1–14. DOI : 10.1007/s00180-015-0635-1. URL : http://dx.doi.org/10.1007/s00180-015-0635-1.

FAVRE-MARTINOZ, Cyril et Thomas MERLY-ALPA (2017). « Constitution et Tirage d'Unités Primaires pour des sondages en mobilisant de l'information spatiale ». *49èmes Journées de statistique de la Société Française de Statistique*.

GANGANATH, Nuwan, Chi-Tsun CHENG et K Tse CHI (2014). « Data clustering with cluster size constraints using a modified k-means algorithm ». *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2014 International Conference on*. IEEE, p. 158–161.

GRAFSTRÖM, Anton (2012). « Spatially correlated Poisson sampling ». *Journal of Statistical Planning and Inference* 142.1, p. 139–147.

GRAFSTRÖM, Anton et J LISIC (2016). « BalancedSampling : Balanced and spatially balanced sampling ». *R package version* 1.2.

GRAFSTRÖM, Anton, Niklas LP LUNDSTRÖM et Lina SCHELIN (2012). « Spatially balanced sampling through the pivotal method ». *Biometrics* 68.2, p. 514–520.

GRAFSTRÖM, Anton et Yves TILLÉ (2013). « Doubly balanced spatial sampling with spreading and restitution of auxiliary totals ». *Environmetrics* 24.2, p. 120–131.

HIDIROGLOU, M.A. et G.B. GRAY (1980). « Algorithm AS 146 : Construction of Joint Probability of Selection for Systematic PPS Sampling ». *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 29.1, p. 107–112.

HOUGH, J Ben et al. (2006). « Determinantal processes and independence ». *Probab. Surv* 3, p. 206–229.

LISIC, Jonathan (2015). « Parcel level agricultural land cover prediction ». Thèse de doct. George Mason University.

LISTER, Andrew J et Charles T SCOTT (2009). « Use of space-filling curves to select sample locations in natural resource monitoring studies ». *Environmental monitoring and assessment* 149.1, p. 71–80.

LOONIS, Vincent (2009). « La construction du nouvel échantillon de l'Enquête Emploi en Continu à partir des fichiers de la Taxe d'Habitation. » *JMS* 2009, p. 23.

LOONIS, Vincent et Xavier MARY (2018). « Determinantal sampling designs ». *Journal of Statistical Planning and Inference*.

MADOW, William G (1949). « On the theory of systematic sampling, II ». *The Annals of Mathematical Statistics*, p. 333–354.

MALINEN, Mikko I et Pasi FRÄNTI (2014). « Balanced k-means for clustering ». *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Springer, p. 32–41.

MARKER, David A et Don L STEVENS (2009). « Sampling and inference in environmental surveys ». *Handbook of Statistics* 29, p. 487–512.

PINCIARO, Susan J (1978). « An algorithm for calculating joint inclusion probabilities under PPS systematic sampling ». *of : ASA Proceedings of the Section on Survey Research Methods*, p. 740–740.

STEVENS JR, Don L et Anthony R OLSEN (2004). « Spatially balanced sampling of natural resources ». *Journal of the American Statistical Association* 99.465, p. 262–278.

TAI, Chen-Ling et Chen-Shu WANG (2017). « Balanced k-Means ». *Asian Conference on Intelligent Information and Database Systems*. Springer, p. 75–82.

11. Économétrie spatiale sur données d'enquête

RAPHAËL LARDEUX, THOMAS MERLY-ALPA
Insee

Résumé

L'économétrie spatiale requiert des données exhaustives sur un territoire, ce qui interdit en principe l'utilisation de données d'enquête. Le présent chapitre présente les écueils liés à l'estimation d'un modèle spatial autorégressif (SAR) sur données échantillonnées et évalue les potentielles corrections proposées par la littérature empirique. Nous identifions deux sources de biais : *(i)* un "effet taille" résultant de la distorsion de la matrice de pondération spatiale et *(ii)* un effet résultant de l'omission d'unités spatialement corrélées avec les unités observées. Tous deux tendent à sous-estimer la corrélation spatiale. Le biais est cependant plus faible dans le cas d'un sondage par grappes et lorsque l'échantillon est suffisamment grand. Deux catégories de méthodes sont proposées par la littérature empirique afin de passer outre ces écueils : l'imputation des valeurs manquantes (régression linéaire, hot deck) et l'agrégation des données à une échelle supérieure. La difficulté est de reconstituer une information complexe à partir de peu d'observations, même si l'imputation par hot deck statistique semble constituer une piste prometteuse. La dernière partie de ce chapitre illustre cette problématique dans le cas concret de l'estimation d'externalités de production entre les industries du département français des Bouches-du-Rhône.

R La lecture préalable des chapitres 2 : "Codifier la structure de voisinage", 3 : "Indices d'autocorrélation spatiale" et 6 : "économétrie spatiale : modèles courants" est nécessaire pour comprendre ce chapitre.

Introduction

Les développement récents de l'économétrie spatiale et de la géolocalisation permettent l'analyse de phénomènes spatiaux à des échelles très locales (firmes, logements, ...), renforçant ainsi la précision des estimations. Les concepts issus de ce champ sont utilisés dans des domaines de plus en plus diversifiés : géostatistique, économie, analyse de réseaux. Cependant, l'application de ces méthodes d'analyse spatiale requiert des données exhaustives, qui ne sont pas toujours accessibles (non-réponse, temps de collecte trop important, ...) et ne peuvent pas aisément être traitées en un temps restreint. L'extension de l'économétrie spatiale aux données d'enquête permettrait de tirer pleinement parti d'une information détaillée pour mesurer finement l'incidence des corrélations spatiales sur les estimations économétriques [1]. Dans ce chapitre, nous discutons ainsi les développements récents relatifs à l'application des méthodes d'estimation spatiales lorsqu'une partie des observations est manquante, en particulier dans le cas de données d'enquête. Nous ne traitons ni la possibilité d'un sondage spatialisé, qui est complexe dans le cas des données sociales [2], ni les cas d'observations dont la localisation est inconnue. Pourquoi l'économétrie spatiale requiert-elle des données exhaustives ? L'économétrie classique repose sur une hypothèse d'indépendance mutuelle des observations. Estimer un modèle sur un sous-ensemble de données peut affecter la puissance des tests statistiques mais, en l'absence de problème de sélection, les estimateurs restent sans biais et efficaces. Au contraire, dans les modèles d'économétrie spatiale, les observations sont considérées comme corrélées entre elles : chaque unité est influencée par ses voisins. Supprimer des observations revient à omettre leurs liens avec les unités observées proches, ce qui introduit un biais dans l'estimation du paramètre de corrélation spatiale et des effets spatiaux estimés. Nous constatons que ce biais tend à atténuer la valeur du paramètre de corrélation spatiale, puisque certains liens de voisinage ne sont alors plus pris en compte dans l'estimation.

Conceptuellement, l'économétrie spatiale se distingue de l'économétrie classique par la façon dont elle considère les observations. En économétrie classique, les observations s'apparentent à un échantillon aléatoire représentatif d'une population et sont interchangeables. L'analyse spatiale les conçoit comme l'unique réalisation d'un processus spatial, chaque observation étant alors nécessaire à l'estimation du processus sous-jacent [3]. L'économétrie spatiale a été développée dans le cadre très pur des modèles de CLIFF et al. 1972, caractérisé par une information exhaustive et parfaite sur les unités spatiales et par l'absence de données manquantes (ARBIA et al. 2016). En pratique, ces conditions ne sont quasiment jamais réunies et appliquer directement des techniques d'estimation spatiale peut fortement altérer les résultats.

L'application de méthodes spatiales à des données non exhaustives pose plusieurs problèmes. Premièrement, les estimations sont perturbées par un "effet taille". L'existence de m données manquantes parmi une population de taille n donne lieu à une matrice de pondération de taille $(n-m)\times(n-m)$ au lieu de la vraie matrice de taille $n\times n$, ce qui biaise le paramètre de corrélation spatiale du simple fait du changement de dimension (ARBIA et al. 2016). Nous illustrons par la suite ce phénomène à partir d'un échantillon localement exhaustif de données simulées par un modèle SAR, en montrant qu'appliquer le même modèle à cet échantillon ne permet pas de retrouver la valeur du paramètre de corrélation spatiale. Deuxièmement, l'existence de données manquantes

1. PINKSE et al. 2010 qualifient ces perspectives de "futur de l'économétrie spatiale".
2. Sur ces questions, on pourra se reporter au chapitre 10 : "échantillonnage spatial".
3. L'analyse spatiale se rapproche en cela des séries temporelles, où le jeu de données observé est issu d'un processus stochastique.

engendre une erreur de mesure sur l'effet du voisinage (régresseur $\mathbf{W}y$) qui biaise les paramètres estimés. Par simulation, nous montrons qu'au-delà de l' "effet taille", ce biais a des conséquences importantes.

Différentes corrections ont été proposées, sans qu'aucune ne s'impose radicalement [4]. Lorsque la localisation des individus est connue, les solutions par imputation sont généralement privilégiées (RUBIN 1976 ; LITTLE 1988 ; LITTLE et al. 2002). Cependant, une imputation naïve, par exemple par un modèle linéaire, ne permet pas de corriger les biais (BELOTTI et al. 2017a). Pour contourner ce problème, KELEJIAN et al. 2010b développent des estimateurs lorsque seul un sous-ensemble incomplet d'une population est disponible. WANG et al. 2013a mettent en place une méthode d'imputation par moindres carrés en deux étapes dans un cadre où des valeurs de la variable dépendante sont aléatoirement manquantes. Dans ce même contexte, LESAGE et al. 2004 recourent à l'algorithme EM (DEMPSTER et al. 1977) : une phase "E" (espérance) assigne une valeur aux données manquantes, conditionnellement aux observables et aux paramètres du modèle spatial sous-jacent, puis une phase "M" (maximisation) détermine la valeur de ces paramètres par maximisation de la vraisemblance du modèle. Par itération, cette procédure permet de tirer d'un modèle estimé l'ensemble de l'information disponible pour imputer des valeurs manquantes. Les travaux plus récents de BOEHMKE et al. 2015 étendent cette procédure au cas d'observations manquantes (variables dépendante et indépendantes inconnues).

Des travaux empiriques récents illustrent l'importance de ces corrections. Dans un modèle de prix hédoniques, LESAGE et al. 2004 appliquent l'algorithme EM pour prédire la valeur des logements non vendus. Dans un modèle de réseaux avec autocorrélation spatiale, LIU et al. 2017 montrent que la détection d'un effet de pair requiert de prendre en compte le processus d'échantillonnage. Les méthodes complexes d'imputation selon un modèle estimé (*model-based*) sont cependant encore peu appliquées. Lorsque certaines données sont manquantes, la solution généralement retenue est de supprimer du champ de l'analyse les observations correspondantes, au risque d'engendrer un biais d'atténuation de la corrélation spatiale. Certains travaux se restreignent à un sous-ensemble, notamment une région ou un groupe particulier (REVELLI et al. 2007), ce qui peut poser un problème d' "effet taille" et amener à sous-estimer les corrélations à la bordure de l'espace considéré (KELEJIAN et al. 2010b). Enfin, la plupart des applications sont réalisées sur données agrégées pour bénéficier de données exhaustives sur une échelle plus large, mais cette solution peut provoquer des erreurs positionnelles [5] (ARBIA et al. 2016) ainsi qu'un biais écologique (ANSELIN 2002b). Nous discutons par la suite l'incidence de ces diverses méthodes sur les estimations spatiales.

Le problème des valeurs manquantes dans un cadre d'observations non indépendantes a été mis en avant par des champs proches de l'économétrie spatiale : séries temporelles et géostatistique d'une part, économétrie des réseaux d'autre part. Les séries temporelles et la géostatistique se rapprochent du traitement des données spatiales continues. Le problème des données manquantes a été abordé très tôt dans le domaine des séries temporelles (CHOW et al. 1976, FERREIRO 1987). JONES 1980 ; HARVEY et al. 1984 recommandent l'utilisation d'un filtre de Kalman pour simultanément estimer un modèle et imputer des valeurs. L'analyse géostatistique corrige des

4. En particulier, ces méthodes varient selon les hypothèses sous-jacentes portant sur les données manquantes : selon que la valeur et/ou la localisation des observations est manquante, que les variables dépendantes et/ou indépendantes sont affectées et selon que la probabilité pour une donnée d'être manquante dépend des corrélations avec les données observables et/ou inobservables. La littérature sur l'incidence des données manquantes établit ainsi une distinction entre *Missing at Random* (MAR), *Missing Completely at Random* (MCAR) et *Missing Not at random* (MNAR). cf RUBIN 1976, HUISMAN 2014

5. ARBIA et al. 2016 proposent ce concept pour désigner les cas où la position d'une observation (X,Y) n'est pas connue précisément. Par exemple, manque de précision dans la mesure, mesure brouillée pour des questions de confidentialité, adresses manquantes.

jeux de données incomplets soit en amont par des méthodes d'échantillonnage spatialisé, soit en prédisant la valeur d'une variable spatiale continue en une position inconnue (interpolation spatiale ou krigeage, voir chapitre 5 :"Géostatistique"). Des approches spatio-temporelles croisant krigeage et filtre de Kalman ont également été développées (MARDIA et al. 1998). Cependant, ces méthodes propres aux données continues ne peuvent être transposées à l'analyse économique et sociale, où les données sont fondamentalement discrètes. De plus, le recours à ces techniques de sondage spatialisé irait à l'encontre des principes fondamentaux de la collecte de données sociales tels que l'équipondération et l'utilisation de bases de sondages déterministes. L'économétrie des réseaux a très vite souligné les biais engendrés par des observations manquantes (BURT 1987 ; STORK et al. 1992 ; KOSSINETS 2006), mais les solutions pratiques restent rares, même si les enjeux liés à l'estimation de l'autocorrélation spatiale sur un échantillon d'un réseau prennent de l'ampleur avec l'utilisation croissante des réseaux sociaux (ZHOU et al. 2017). De même qu'en économétrie spatiale, la principale difficulté est de reconstituer l'information sur les données inobservées à partir des données observées, sans connaître l'effet des premières sur les secondes (KOSKINEN et al. 2010). En particulier, HUISMAN 2014 ne tranche pas entre diverses stratégies d'imputation classiques et montre que celles-ci ne fonctionnent que dans des cas spécifiques. Des solutions fondées sur des méthodes d'échantillonnage ont également été proposées afin de collecter des données sur les populations d'intérêt (GILE et al. 2010).

Le présent chapitre se concentre sur deux questions : quels sont les biais engendrés par l'application de méthodes spatialisées à des données d'enquête ? quelles sont les conséquences des diverses solutions classiques (suppression des données, imputation, agrégation) ? Ces questions sont abordées par ARBIA et al. 2016, qui procèdent par simulation et observent une incidence plus marquée des données manquantes lorsqu'elles sont regroupées en grappes, auquel cas l'intégralité de phénomènes locaux peut être perdue. Ils considèrent cependant des cas où les données manquantes représentent au maximum 25 % de la population, ce qui est très faible par rapport aux données d'enquête, où elles atteignent généralement plus de 90 % de la population.

La section 11.1 présente les biais issus de l'application de méthodes spatiales à un échantillon non exhaustif de données, selon la part des observations échantillonnées et le type de sondage. La section 11.2 discute les conséquences de quelques solutions usuelles : le passage à l'échelle supérieure par agrégation et l'imputation des valeurs manquantes. La section 11.3 illustre ces biais à partir de l'estimation d'une équation de production avec externalités sur les industries du département français des Bouches-du-Rhône.

11.1 Première approche par simulations

Dans cette première partie, nous mettons en évidence, par des simulations de type Monte-Carlo, l'existence d'un biais dans l'estimation d'un modèle autorégressif spatial (SAR) sur des données d'échantillon. D'abord, nous simulons sur un espace géographique des données spatialement corrélées en fixant la valeur du paramètre de corrélation spatiale. Nous procédons par la suite à des tirages d'échantillon, à partir desquels nous estimons la valeur du paramètre de corrélation spatiale.

11.1.1 Simulation d'un SAR

L'espace géographique retenu est une carte de l'Europe [6], détaillée au niveau administratif NUTS3 (échelon le plus bas dans la hiérarchie NUTS définie par Eurostat, qui correspond à des petites zones sur lesquelles peuvent être menées des études spécifiques : les départements français, par exemple) de laquelle nous retirons les îles les plus éloignées ainsi que l'Islande afin de conserver

6. Cette carte est diffusée sur le site : `http://ec.europa.eu/eurostat/fr/web/gisco/geodata/reference-data/administrative-units-statistical-units`.

un espace géographique homogène et compact. À partir du *shapefile* de l'Europe, nous construisons une matrice de voisinage $\mathbf{W}$ fondée sur la distance, de telle sorte que le poids associé à deux unités voisines décroît selon le carré de la distance et s'annule lorsque cette distance dépasse un seuil limite. Les résidus et une variable explicative sont simulés dans des lois normales : $\varepsilon \sim \mathcal{N}(0,1)$ et $X \sim \mathcal{N}(5,2)$. Cela nous permet de finalement simuler une variable Y suivant un modèle de type SAR (*Spatial Auto-Regressive*) :

$$Y = (1-\rho\mathbf{W})^{-1} X\beta + (1-\rho\mathbf{W})^{-1}\varepsilon \tag{11.1}$$

avec $\beta = 1$ et $\rho = 0.5$, paramètres de référence que nous cherchons à retrouver par l'estimation de modèles SAR sur des échantillons. Les données des variables simulées Y sont représentées sur la figure 11.1. La présence de zones colorées concentrées est caractéristique de l'autocorrélation spatiale positive résultant du processus générateur des données.

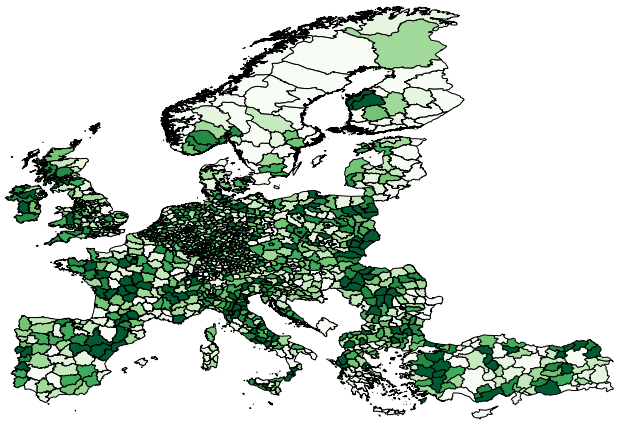


FIGURE 11.1 – Y simulé selon un modèle SAR
Copyright : EuroGeographics pour les limites administratives

La table 11.1 présente les résultats de l'estimation d'un modèle SAR sur l'ensemble des zones NUTS3 d'Europe. Ils confirment la validité de cette simulation, puisque les paramètres β et ρ estimés sont très proches des valeurs calibrées initialement.

β	ρ	Direct	Indirect	Total
0.989	0.494	1.043	0.860	1.902

TABLE 11.1 – Paramètres estimés par SAR sur l'ensemble des zones

Encadré 11.1.1 — Simulation d'un SAR avec R. Pour simuler un SAR en R, l'étape la plus importante est de formater sa matrice de voisinage $\mathbf{W}$ de la façon suivante :

```
D <- nb2listw(W, style="W", zero.policy=TRUE)
```

Une fois la matrice de voisinage au format `listw`, il faut alors inverser $1 - \rho\mathbf{W}$ en utilisant la fonction suivante, dont ρ est l'un des paramètres :

```
InvD <- invIrW(D,rho)
```

Attention, cette étape peut être chronophage. Il ne reste alors plus qu'à simuler notre variable Y :

```
Y <- (InvD %*% X) + (InvD %*% eps)
```

11.1.2 Procédures de sondage

L'enjeu est d'examiner la capacité des modèles spatiaux à correctement estimer ρ et β à partir d'échantillons tirés dans ces données simulées. Nous discutons en particulier l'effet que peut avoir l'échantillonnage de certaines de ces zones sur l'estimation du modèle sous-jacent.

Un sondage consiste à sélectionner de façon aléatoire en suivant une procédure dite plan d'échantillonnage un ensemble de n unités au sein d'une population de N, où n est souvent bien plus petit que N afin de limiter les coûts liés à la collecte d'information. La théorie des sondages affirme que les estimations réalisées à l'aide de l'échantillon s'étendent sans biais à la population totale, mais que celles-ci sont plus précises quand la taille de l'échantillon augmente et quand le plan d'échantillonnage est adapté à la variable estimée. Pour approfondir les questions de sondage, la lecture de ARDILLY 1994, TILLÉ 2001 ou COCHRAN 2007 est conseillée.

Dans la suite de cette partie, nous présentons quelques techniques de sondage classiques et leur application dans le cadre des NUTS3 européens. Nous pouvons cependant déjà faire quelques hypothèses et remarques générales, en suivant les idées développées dans GOULARD et al. 2013 concernant le nouveau recensement de la population. D'une part, l'effet ne devrait évidemment pas être le même selon la taille n de l'échantillon retenu. Avec une petite dizaine de zones, la structure spatiale initiale ne pourra pas être reconstituée, tandis qu'échantillonner 95 % voire 99 % des zones devrait permettre de la retrouver facilement. D'autre part, la question de la méthode d'échantillonnage va également se poser : la dimension spatiale est-elle prise en compte dans le cadre de la méthode ? On pourra se reporter au chapitre 10 : "échantillonnage spatial" pour approfondir ces questions.

Sondage aléatoire simple

Le sondage aléatoire simple consiste à tirer indépendamment et sans remise n boules au sein d'une urne de taille N. Les individus ont alors tous la même chance d'être sélectionnés dans l'échantillon. La sélection d'un individu dans un échantillon diminue la probabilité qu'ont les autres d'être également inclus. Dans notre cas, on sélectionne n zones complètement au hasard. La figure 11.2 présente un exemple d'échantillon.

Sondage poissonnien

Le sondage poissonnien (ou bernoullien) consiste à tirer à pile ou face pour chaque individu de la population son appartenance à l'échantillon. Dans ce cas, les individus ont toujours la même chance d'être sélectionnés dans l'échantillon. La sélection d'un individu dans un échantillon n'influe pas sur la probabilité qu'ont les autres d'être également inclus, mais la taille de l'échantillon n'est pas fixée *a priori*. Dans notre cas, chaque zone a une probabilité p d'être retenue dans l'échantillon : la taille de l'échantillon obtenu est alors pN en espérance.

Sondage par grappes

Le sondage par grappes (ou aréolaire) consiste à sélectionner ensemble des groupes d'individus. Les individus ont toujours la même chance d'être sélectionnés dans l'échantillon. Cependant, la sélection d'un individu au sein d'un échantillon influe fortement sur la probabilité qu'ont les autres d'être également inclus, car les individus d'une même grappe sont toujours sélectionnés ensemble.

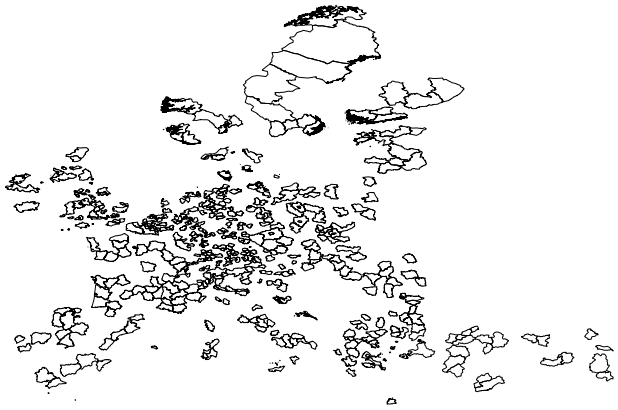


FIGURE 11.2 – Un échantillon obtenu par sondage aléatoire simple ($n = 500$)
Copyright : EuroGeographics pour les limites administratives

Ici, il s'agit de rassembler les zones NUTS3 en différentes grappes et ensuite de réaliser une sélection aléatoire de certaines de ces grappes. L'intérêt principal est de limiter les coûts de collecte, au prix d'une perte en précision liée à l'homogénéité intra-grappe.

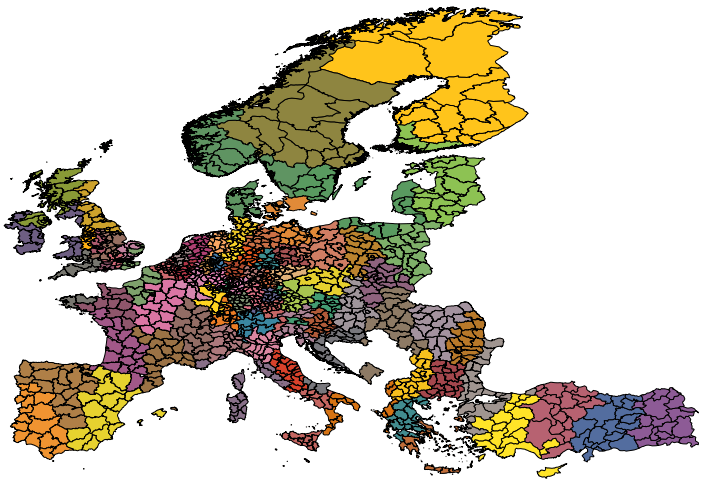


FIGURE 11.3 – Division de l'Europe en grappes
Copyright : EuroGeographics pour les limites administratives

Il serait envisageable d'utiliser les différents niveaux NUTS1 ou NUTS2 comme grappes. Cependant, ils sont de taille importante et ne comportent pas tous le même nombre de NUTS3. Or des grappes de taille trop importante vont limiter le nombre de simulations possibles. Au contraire, des grappes comportant des nombres de zones différents introduisent soit une problématique

de poids de sondage différents entre les individus, que nous ne souhaitons pas traiter ici (voir DAVEZIES et al. 2009 pour un débat sur l'usage des poids de sondage en économétrie), soit une problématique de taille d'échantillon variable ce qui peut avoir des effets complexes à analyser. Nous séparons donc les zones en grappes de même taille tout en maintenant une certaine cohérence géographique. Comme la matrice de pondération est basée sur la distance géographique, nous privilégions les grappes les moins étendues possible.

Afin d'obtenir des grappes de taille identique, il est nécessaire que le nombre de grappes soit un diviseur du nombre de zones NUTS3. En vue de limiter la taille des grappes, nous rassemblons les 1445 zones NUTS3 en 85 grappes de 17 zones chacune. Pour cela, nous utilisons un algorithme de construction des grappes : partant de la zone la plus éloignée du centre de la carte, nous agrégeons les zones les plus proches de celle-ci jusqu'à en obtenir 17. Comme les grappes sont construites une à une, les NUTS3 les plus éloignés seront déjà affectés pour la construction des grappes précédentes, et l'algorithme se poursuit avec des zones plus centrales. Les grappes obtenues sont représentées sur la figure 11.3.

Sondage stratifié

Le sondage stratifié correspond à un tirage de n boules, mais en tirant n_1 boules dans une première urne, n_2 dans une deuxième, jusqu'à n_H dans une H-ième, où $n = n_1 + n_2 + \ldots + n_H$. Pour réaliser un tirage stratifié, il convient de bien définir les H strates d'une part, et de bien choisir l'allocation $(n_1, \ldots, n_H)$ d'autre part. Une allocation classique est l'allocation de Neyman, qui a pour propriété de minimiser la variance de l'estimateur du total d'une variable d'intérêt (voir par exemple TILLÉ 2001). La formule est la suivante :

$$n_h = n \frac{N_h S_h}{\sum_{i=1}^{H} N_i S_i} \tag{11.2}$$

avec n la taille de l'échantillon total, N_h la taille de la strate h et S_h la dispersion de la variable d'intérêt au sein de la strate h. Dans certains cas, lorsque les comportements vis-à-vis de la variable d'intérêt sont hétérogènes, cette formule peut conduire à enquêter exhaustivement certaines strates, c'est-à-dire à leur appliquer un taux de sondage de 100 %.

11.1.3 Résultats et interprétation

Afin d'estimer l'effet de l'échantillonnage d'une partie des NUTS3 européens, nous suivons la méthode de Monte-Carlo. Nous réalisons ainsi 100 simulations de Y selon un modèle SAR puis nous tirons 100 échantillons pour chacune d'entre elles. Le modèle SAR est estimé sur chaque échantillon et l'on récupère les paramètres d'intérêt. Enfin, les paramètres présentés en résultat sont les moyennes des ρ et β sur les 10 000 échantillons et leurs écart-types sont calculés sur ces 10 000 valeurs.

Pour chacun des 10 000 tirages, sont ainsi conservées les valeurs de X et de Y des zones échantillonées. On reconstruit alors une matrice de pondération spatiale $\mathbf{W}_{\text{echantillon}}$ fondée sur la distance, tel que précédemment, mais limitée aux unités présentes dans l'échantillon. Différentes tailles d'échantillon et différentes méthodes d'échantillonnage sont considérées.

Sondage aléatoire simple

La table 11.2 présente les estimations obtenues dans le cas d'un sondage aléatoire simple pour des tailles d'échantillon n variant de 50 à 250 zones. Ces estimations mettent en évidence une autocorrélation spatiale significative à partir d'un échantillon de taille $n = 150$, ce qui correspond approximativement, dans notre cas, à un taux de sondage de $1/10$. Le paramètre β est estimé sans biais quelle que soit la taille de l'échantillon, mais le paramètre estimé $\hat{\rho}$ est nettement inférieur

à sa vraie valeur $\rho = 0.5$. Par conséquent, pour des échantillons de petite taille, l'effet indirect n'est pas significativement différent de zéro et reste bien inférieur à celui observé sur la population entière. L'autocorrélation spatiale est largement sous-estimée.

n	$\hat{\rho}$	$\hat{\beta}$	Direct	Indirect	Total
50	0.043	1.055***	1.056***	0.016	1.072***
	(0.043)	(0.125)	(0.125)	(0.017)	(0.128)
100	0.058*	1.050***	1.052***	0.032*	1.083***
	(0.031)	(0.087)	(0.087)	(0.019)	(0.091)
150	0.072**	1.049***	1.051***	0.048**	1.099***
	(0.028)	(0.068)	(0.068)	(0.020)	(0.073)
250	0.101***	1.051***	1.054***	0.080***	1.135***
	(0.026)	(0.051)	(0.052)	(0.023)	(0.060)

TABLE 11.2 – Estimation d'un modèle SAR sur des échantillons tirés par sondage aléatoire simple
Note : *** désigne une significativité à 1 %, ** une significativité à 5 % et * une significativité à 10 %. Les écart-types sont entre parenthèses. n : nombre d'observation dans l'échantillon. Ces estimations proviennent de 10 000 simulations.

Sondage par grappes

L'échantillonnage par grappes permet de conserver une structure géographique forte localement, ce qui dans notre cas semble bénéfique pour la détection d'effets spatiaux, en particulier pour des petites valeurs de n. À partir des grappes présentées dans la partie 11.1.2, nous réalisons des tirages de nombres différents de grappes, allant de 3 à 15 grappes, soit 51 à 255 zones. La table 11.3 montre les résultats obtenus pour des valeurs de $n = 17p$, la taille de l'échantillon composé de p grappes.

n	p	$\hat{\rho}$	$\hat{\beta}$	Direct	Indirect	Total
51	3	0.309*	1.015***	1.051***	0.441*	1.492***
		(0.237)	(0.091)	(0.097)	(0.262)	(0.310)
102	6	0.348***	1.017***	1.054***	0.493***	1.546***
		(0.100)	(0.063)	(0.066)	(0.188)	(0.215)
153	9	0.363***	1.017***	1.054***	0.516***	1.571***
		(0.078)	(0.052)	(0.055)	(0.152)	(0.176)
255	15	0.377***	1.014***	1.052***	0.541***	1.593***
		(0.058)	(0.038)	(0.040)	(0.119)	(0.136)

TABLE 11.3 – Estimation d'un modèle SAR sur des échantillons tirés par grappes
Note : *** désigne une significativité à 1 %, ** une significativité à 5 % et * une significativité à 10 %. Les écart-types sont entre parenthèses. n : nombre d'observations dans l'échantillon. p : nombre de grappes dans l'échantillon. Ces estimations proviennent de 10 000 simulations.

Avec un sondage par grappes, le paramètre $\hat{\rho}$ est plus proche de sa vraie valeur et l'inclut dans son intervalle de confiance. La précision de l'estimation s'améliore nettement lorsque n augmente, mais l'estimateur reste biaisé. Ainsi, contrairement au cas du sondage aléatoire simple, il est possible de capter les interactions spatiales même avec un taux de sondage très faible de l'ordre de 3 %. En effet, les unités enquêtées sont fortement concentrées dans l'espace et donc très représentatives de la corrélation spatiale. En revanche, si le nombre d'unités tirées est faible, il en va de même pour la précision de l'estimation de la corrélation spatiale. Dès lors, l'effet indirect

est bien détecté même pour des échantillons de petite taille et sa valeur est plus proche de celle obtenue sur la population totale. L'estimation d'effets géographiques semble ainsi raisonnable dans le cadre d'un tel type de sondage.

Deux questions subsistent : tout d'abord, est-ce que cet échantillonnage par grappes n'aurait pas tendance à favoriser la détection d'un modèle autocorrélé spatialement, même si la tendance n'est pas majeure sur la totalité de la population ? Comme l'on dispose de peu de valeurs de X et Y, le terme $\mathbf{W}Y$ est paradoxalement assez bien connu, ce qui pourrait amener à favoriser cette piste. D'autre part, et cela sera developpé dans la partie 11.1.4, on peut s'étonner de l'écart observé entre le $\hat{\rho}$ estimé et la vraie valeur utilisée pour la génération du SAR, alors même qu'on détecte bien les effets spatiaux.

11.1.4 Un "effet taille"

Les résultats obtenus par simulation peuvent étonner les économètres. En effet, dans le cadre d'un sondage aléatoire simple, assimilable au modèle de superpopulation utilisé en économétrie [7], l'estimation d'un paramètre d'une population ou d'un modèle est usuellement sans biais, tant que le plan d'échantillonnage est correctement spécifié. Il apparaît alors que le paramètre d'autocorrélation spatiale ρ ne suit pas cette "loi" classique de la théorie des sondages [8].

Nonobstant la question de la sélection aléatoire des zones sur lesquelles on récupère l'information relative aux Y, se restreindre à un nombre de zones inférieur à celui de la population entière induit une modification de la structure spatiale sous-jacente.

La question des biais écologiques, c'est-à-dire des erreurs d'estimation de modèles économétriques spatiaux qui proviennent de la mauvaise spécification spatiale, que cela soit au niveau du grain (de la résolution) des données ou des problèmes frontaliers, est proche de ce sujet. Ainsi, il est tout à fait possible que lorsque que l'on se restreint à n zones, avec $n < N$, on ne puisse jamais obtenir un effet spatial aussi fort que sur la population entière.

Pour illustrer ce point, nous estimons plusieurs centaines de modèles SAR, générés sur la population entière, à partir d'une restriction de la population aux n NUTS3 les plus centraux : l'idée étant de se limiter à une sous-partie de l'Europe sans qu'elle ne soit choisie aléatoirement ni de façon morcelée comme c'était le cas pour les échantillons obtenus précédemment (figure 11.2). La figure 11.4 compare les valeurs de $\hat{\rho}$ obtenues en suivant trois protocoles pour différents pourcentages P % de la population totale : la sélection des P % NUTS3 les plus centraux, un sondage par grappes où chaque grappe de zones a P % de chances d'être sélectionnée, et un sondage poissonnien classique où chaque zone a indépendamment P % de chances d'être sélectionnée.

De même que dans la partie 11.1.3, le sondage poissonnien (proche du sondage aléatoire simple) donne des valeurs estimées de $\hat{\rho}$ bien plus faibles que le sondage par grappes. L'apport principal de cette figure est dans la courbe rouge, qui repose sur une sélection non aléatoire d'une partie des zones. Elle converge plus rapidement que les autres vers 0.5, la vraie valeur de ρ. Ce constat semble confirmer l'hypothèse d'un biais lié à la déformation de la structure spatiale ou "effet taille", résultant d'une restriction à un sous-ensemble de la population totale.

7. Ce terme est lié à la différence entre approche *sous le plan* et *sous le modèle* en sondages. Si on raisonne sous le plan, on suppose que la population a des valeurs de Y déterministes - approche usuelle. Sous le modèle, on suppose qu'il y a un modèle dit de *superpopulation* dont dérivent les Y de la population. Ici on doit suivre cette approche pour pouvoir estimer nos modèles SAR

8. On notera que plusieurs paramètres ne respectent pas cette loi : on peut par exemple penser au maximum d'une variable Y sur une population, qui n'est pas possible à estimer sans biais à partir d'un échantillon. Par ailleurs, dans notre cas de sondage aléatoire simple ou par grappes, il n'y a pas de problèmes de sous-couverture, c'est-à-dire d'unités de la population qui ne peuvent pas appartenir à l'échantillon pour des raisons souvent liées à la qualité des registres. Cette piste ne peut pas expliquer le biais sur $\hat{\rho}$.

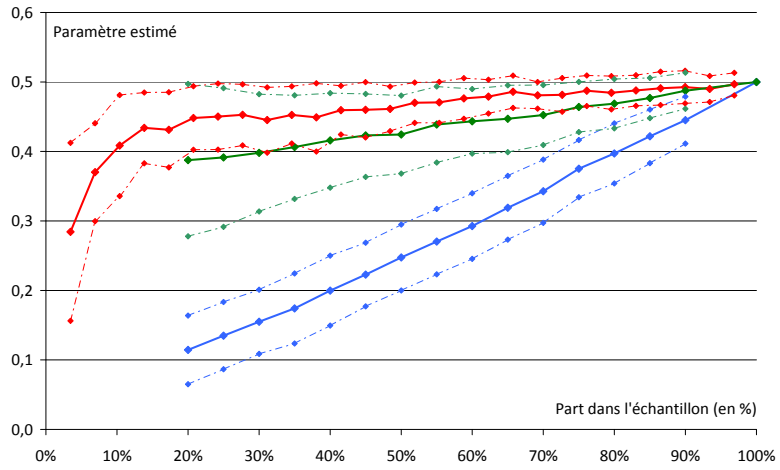


FIGURE 11.4 – L' "effet taille"
Note : Chaque point d'une courbe en trait plein représente une estimation du paramètre $\hat{\rho}$ pour une taille d'échantillon exprimée en pourcentage de la population exhaustive. Lorsque l'estimation est réalisée sur données exhaustives, on retrouve $\hat{\rho} = 0.5$. La courbe bleue correspond à un sondage aléatoire simple, la courbe verte à un tirage par grappes. La courbe rouge représente un sélection déterministe des régions, en partant d'un point initial puis en s'en éloignant progressivement. Le courbes en pointillés représentent les intervalles de confiance à 95 %

11.1.5 Robustesse des résultats

En conclusion de cette section, notons que la spécification retenue pour le modèle spatial n'affecte les résultats obtenus que de façon marginale. Ces derniers restent inchangés lorsque le seuil maximum de distance varie ou lorsque la notion de distance retenue est fondée sur les plus proches voisins (table 11.10 en annexe 11.3.6). Enfin la vraie valeur du paramètre ρ n'affecte pas les résultats des estimations. La figure 11.5 montre que, à taux de sondage donné, une estimation sur un échantillon tiré par sondage aléatoire simple ne permet presque jamais de retrouver la vraie valeur du paramètre ρ. Dans le cas d'un sondage par grappe, cette valeur peut être incluse dans l'intervalle de confiance du paramètre estimé, mais le biais lié à l'estimation ne disparaît pas lorsque son ampleur ou son signe changent. Dans tous les cas, le biais atténue l'ampleur de la corrélation spatiale estimée.

Enfin, considérer un modèle de type SEM (*Spatial Error Model*) : $Y_2 = X\beta + (1 - \lambda \mathbf{W})^{-1}\varepsilon$ n'affecte pas radicalement les résultats (table 11.11 en annexe 11.3.6).

11.2 Pistes de résolution

Une première position face à un problème de données manquantes est d'ignorer, consciemment ou non, ces données et d'appliquer directement le modèle spatial aux unités observées. Ce choix atténue le paramètre de corrélation spatiale relativement à sa vraie valeur, de par un "effet taille" et un effet d'échantillonnage.

Le premier effet provient de différences entre le modèle théorique et le modèle estimé concernant la dimension de la matrice de pondération spatiale. Pour le supprimer, il faut comparer données exhaustives et données échantillonnées selon une même structure géographique, et donc sur un même nombre d'unités. Pour compenser le second, il faut être en mesure de reconstituer les corrélations spatiales entre unités observées et manquantes. Dans le cas présent, la localisation des unités

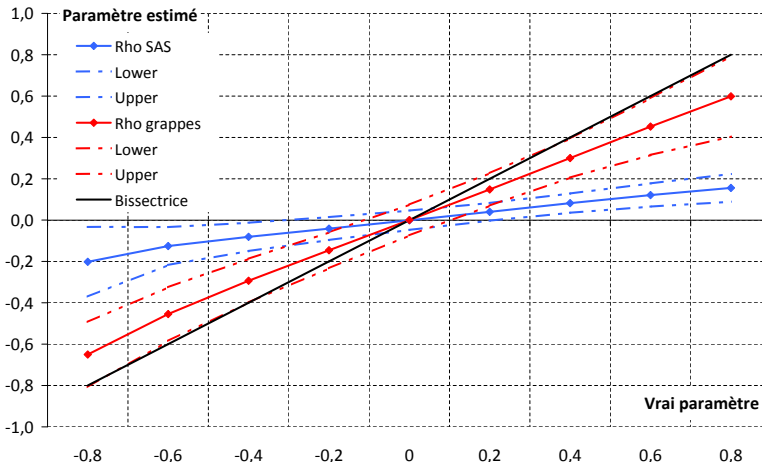


FIGURE 11.5 – Estimations de $\hat{\rho}$ pour diverses valeurs de ρ
Note : Les courbes en trait plein représentent la valeur estimée $\hat{\rho}$ en fonction de la valeur ρ fixée pour la simulation des données. La courbe bleue correspond au cas de données tirées par sondage aléatoire simple et la courbe rouge au cas d'un sondage par grappes. Les courbes en pointillés représentent les intervalles de confiance à 95 % de l'estimateur $\hat{\rho}$

est toujours supposée connue [9].

Dans cette partie, nous discutons l'incidence de deux solutions généralement appliquées dans les travaux empiriques : le passage à une échelle supérieure par agrégation des données et l'imputation des données manquantes. Ces méthodes conservent la structure géographique mais sont plus ou moins efficaces pour reconstituer les corrélations spatiales.

11.2.1 Passer à l'échelle supérieure par agrégation

Problématique

En l'absence de données individuelles exhaustives, de nombreux travaux sont réalisés à une échelle agrégée de régions, de départements, de zones d'emploi. Ce choix dépend de façon cruciale de l'échelle d'analyse de la problématique, requiert de disposer d'un bon estimateur de la moyenne locale et peut mener à des biais écologiques (voir ANSELIN 2002b pour plus de précisions). Les corrélations intra-zone sont alors omises, au profit des corrélations entre zones.

Pour évaluer cette solution, nous simulons 6 000 points selon une loi uniforme sur un espace carré et leur affectons, comme dans la partie 11.1, des valeurs de X et de Y correspondant à un SAR de paramètre d'autocorrélation spatiale $\rho = 0.5$. Ces points sont représentés sur la surface de gauche de la figure 11.6. Puis cet espace carré est découpé selon une grille de taille $G \times G$ pour différentes valeurs de G, et à chaque centroïde de chaque case est affectée la moyenne des points situés à l'intérieur de cette case. Les panneaux du centre et de droite de la figure 11.6 représentent cette configuration pour $G = 50$ et $G = 20$ respectivement.

L'estimation d'un modèle SAR sur données exhaustives agrégées avec $G = 50$ permet d'estimer un paramètre de corrélation spatiale $\hat{\rho} = 0,47$ d'écart-type $\hat{\sigma}_\rho = 0,068$. Ce paramètre est significativement positif et l'estimation inclut $0,5$ dans son intervalle de confiance. L'agrégation de données sur des parcelles limiterait la perte d'interactions spatiales et minimiserait le biais dans l'estimation du paramètre de corrélation spatiale. La figure 11.7 montre que les paramètres ρ et β sont précisément estimés et sans biais dès lors que la grille sur laquelle les données sont agrégées

9. Le manque d'information concernant la localisation de certaines unités est un autre enjeu des recherches actuelles en économétrie spatiale (ARBIA et al. 2016) qui dépasse cependant le cadre du présent chapitre.

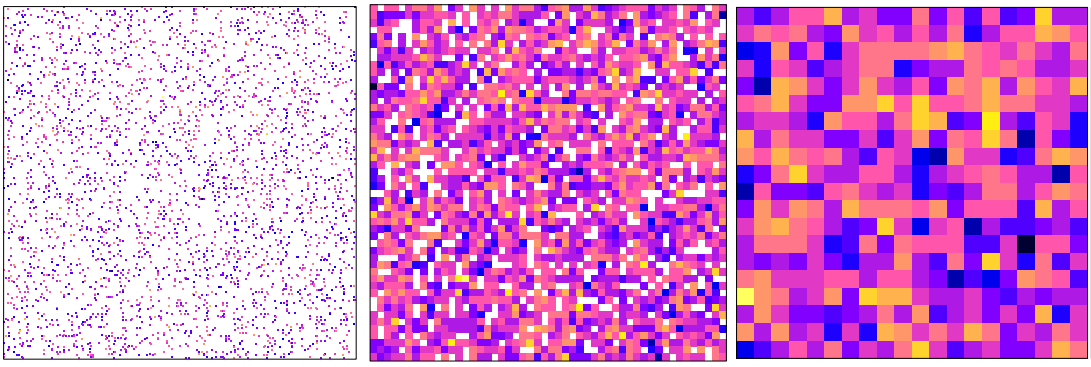


FIGURE 11.6 – Agrégation de données spatiales
Note : *à gauche :* 6 000 points simulés selon une loi uniforme. *au centre :* données agrégées selon 50×50 parcelles. *à droite :* données agrégées selon 20×20 parcelles

est relativement fine. Ainsi, pour des données exhaustives, plus le quadrillage est fin, plus on se rapproche de la structure spatiale des données ponctuelles et, par conséquent, plus la corrélation spatiale estimée est proche de sa vraie valeur.

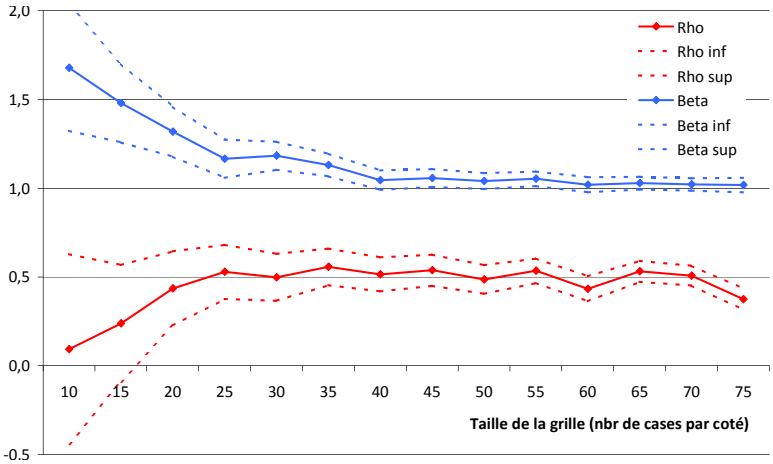


FIGURE 11.7 – Effet de la finesse de la grille sur les paramètres estimés
Note : Les résultats sont obtenus à partir de la simulation de 6 000 points selon une loi uniforme

Application à un échantillon

Cette procédure est répliquée sur des données échantillonnées par sondage aléatoire simple. La finesse de la grille répond à un arbitrage biais-variance : des maillons fins sont plus fidèles aux distances entre observations mais mènent à des estimations de moyennes locales moins précises pour chaque variable. Sous réserve d'assigner des poids nuls ainsi que des valeurs nulles des variables expliquée et explicatives aux mailles sans observation, il est possible de retrouver l'effet spatial simulé.

La table 11.4 présente les résultats de cette procédure pour différentes tailles d'échantillon et diverses grilles spatiales. Dans la majorité des cas, la vraie valeur de ρ se situe bien dans l'intervalle de confiance du paramètre estimé. Pour un petit échantillon, une grille trop grossière écrase les effets spatiaux tandis qu'une grille trop fine fournit une mauvaise estimation des variables individuelles. Comme précédemment, plus l'échantillon est grand, plus l'estimation est précise.

Ces simulations tendent à valider statistiquement l'approche par agrégation, sous réserve que

	$\hat{\rho}$				$\hat{\beta}$			
n \ G	10	30	50	60	10	30	50	60
100	0.487***	0.494***	0.483***	0.478***	1.016***	1.007***	1.027***	1.030***
	(0.070)	(0.060)	(0.068)	(0.072)	(0.134)	(0.115)	(0.129)	(0.134)
200	0.482***	0.499***	0.495***	0.489***	1.020***	0.998***	1.006***	1.011***
	(0.064)	(0.045)	(0.046)	(0.050)	(0.126)	(0.084)	(0.088)	(0.095)
500	-0.093	0.488***	0.483***	0.489***	1.035***	1.022***	1.031***	1.021***
	(0.701)	(0.032)	(0.030)	(0.032)	(0.121)	(0.060)	(0.055)	(0.059)
1000	-0.982	0.487***	0.485***	0.491***	1.048***	1.024***	1.028***	1.019***
	(0.159)	(0.024)	(0.020)	(0.021)	(0.119)	(0.045)	(0.038)	(0.040)

TABLE 11.4 – Estimation d'un SAR sur échantillons agrégés par parcelle
Note : Chaque ligne correspond à une taille d'échantillon n tiré parmi les 6 000 points simulés et chaque colonne correspond à la finesse de la grille en termes de nombre de carreaux (une grille de taille 30 découpe le carré initial en 900 cases) *** désigne une significativité à 1 %.

l'interprétation ne soit pas effectuée directement à l'échelle individuelle, mais reposent cependant sur des hypothèses fortes (coordonnées des unités déterminées par une loi uniforme, processus SAR homogène), rarement vérifiées en pratique.

11.2.2 Imputer les données manquantes

Pour rester à l'échelle des données disponibles, la solution est d'imputer des valeurs aux observations manquantes. C'est là encore une manière de faire abstraction de l' "effet taille" en assurant une cohérence entre la structure spatiale des données d'enquête et des données administratives. Face à des valeurs manquantes dans le cadre d'une enquête ou d'un recensement, attribuer des valeurs "plausibles" à ces unités permet de disposer d'un échantillon voire d'une population complète.

Méthodes d'imputation

Cette partie recense quelques méthodes classiques d'imputation. Le lecteur intéressé pourra se reporter à un bon livre de théorie des sondages, par exemple ARDILLY 1994 ou TILLÉ 2001, pour plus d'informations, de contexte théorique ainsi que pour d'autres méthodes plus avancées. Dans la cas d'une imputation par le ratio ou par hot deck, les variables explicatives X sont supposées connues de façon exhaustive.

Imputation par la moyenne. La méthode d'imputation par la moyenne (ou par la médiane, ou par la classe dominante dans le cas de variables qualitatives) est une méthode usuelle qui consiste à remplacer toutes les valeurs manquantes par la moyenne des valeurs observées. Cette méthode ne respecte pas une éventuelle structure économétrique entre différentes variables de l'enquête et peut conduire à des résultats faux dans le cadre d'estimations de tels modèles.

Imputation par le ratio. La méthode d'imputation par le ratio consiste à mobiliser l'information auxiliaire X disponible sur la totalité de la population, y compris les unités pour lesquelles l'information d'intérêt Y est manquante, afin d'imputer des valeurs de Y plausibles. Pour cela, on postule l'existence d'un modèle linéaire de la forme $Y = \beta X + \varepsilon$. $\hat{\beta}$ est estimé par les moindres carrés ordinaires, puis la valeur $Y_{\text{ratio}} = \hat{\beta} X$ est imputée pour les Y manquants. Le ratio des Y sur les X, dans le cas de données quantitatives, est le même entre les unités observées et les unités pour lesquelles on ne dispose pas d'information. Cette méthode peut être affinée en rajoutant des contraintes sur les unités pour lesquelles on calcule l'estimation du β, par exemple sur un domaine ou sur une strate précise.

Imputation par hot deck. La méthode de hot deck associe un donneur à une valeur manquante

de façon aléatoire, par opposition au cold deck qui établit ce lien de manière déterministe. Un donneur est ici un individu statistiquement "proche" de l'individu manquant (il partage des valeurs proches des X auxiliaires, appartient à la même strate, au même domaine, ou encore se situe à la même position spatiale). La mise en pratique d'un hot deck repose sur la définition d'un critère de distance, à partir duquel sont déterminés k voisins de l'individu dépourvu de valeur Y. Un individu parmi ces k voisins est choisi au hasard, uniformément ou non, pour donner sa valeur pour le nouveau Y_{hotdeck}. Il est possible d'introduire des variantes, par exemple en limitant le nombre de fois où un même individu peut être donneur, ou en réalisant le hot deck de façon séquentielle.

Pour aller plus loin

Les méthodes d'imputation peuvent plus ou moins directement altérer les estimations effectuées sur les données imputées. Le lien entre Y et X sur lequel repose l'imputation peut se retrouver exacerbé dans l'estimation du modèle sur Y et X (voir CHARREAUX et al. 2016 pour une discussion de ce point). De façon similaire, voire même amplifiée, l'utilisation de méthodes d'économétrie spatiale sur de telles données requiert une extrême précaution. En effet, la méthode d'imputation peut faire émerger une structure spatiale *ex-nihilo* ou au contraire briser les corrélations spatiales qu'elle ne prend pas en compte. Des exemples d'application de ces méthodes sont présentés en partie 11.3.6.

Enfin, tel que mentionné en introduction, des méthodes plus raffinées d'imputation au moyen de l'algorithme EM ont été développées (LESAGE et al. 2004 ; WANG et al. 2013a). Elles sont cependant complexes, très spécifiques au type d'information manquante et restent encore peu appliquées.

11.3 Application empirique : la production industrielle dans les Bouches-du-Rhône

Afin d'illustrer les enjeux relatifs à l'estimation de modèles spatiaux sur données d'enquête, nous procédons dans cette section à l'estimation d'une fonction de production sur des établissements issus du répertoire SIRUS. L'approche spatiale permet de mesurer l'incidence des interactions entre les processus de production des diverses entreprises sur la production de chacune d'entre elles. De tels *spillovers* entre entreprises ont déjà été mis en évidence par une importante littérature sur les économies d'agglomération, voir notamment LESAGE et al. 2007, ERTUR et al. 2007, LÓPEZ-BAZO et al. 2004, EGGER et al. 2006.

11.3.1 Données

Le répertoire SIRUS (Système d'Identification au Répertoire des Unités Statistiques) est le répertoire référent en termes de champ de la statistique d'entreprises française. Il est composé des entreprises, des groupes et de leurs établissements, contenus dans SIRENE (Système Informatisé du RÉpertoire National des Entreprises et des établissements), le répertoire administratif qui permet l'enregistrement des unités légales. Sont enregistrés pour chaque entreprise son chiffre d'affaires, son activité principale (disponible *via* le code APE, suivant la nomenclature française), son total de bilan, ses exportations, son effectif (tant administratif qu'en équivalent temps plein), son adresse physique ainsi que la liste des établissements qui la composent.

Les informations géographiques disponibles sur les établissements ont permis, grâce à un travail réalisé par la Division des Méthodes et Référentiels Géographiques de l'Insee, de géolocaliser aux coordonnées (x,y) chacun d'entre eux. Pour cela, différentes données ont été utilisées : du plus précis au moins précis, la référence cadastrale, la voie puis le centre de la commune pour les cas pour lesquels on dispose de trop peu d'information. Ces données géographiques, associées aux données économiques disponibles dans le répertoire SIRUS, permettent la modélisation de relations économétriques en prenant en compte la structure spatiale.

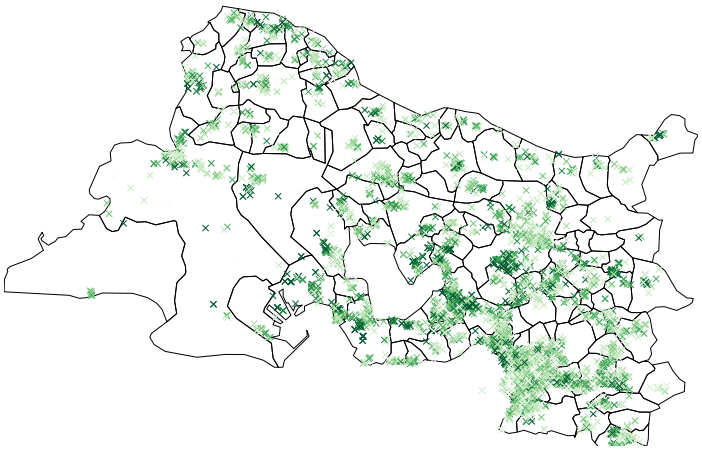


FIGURE 11.8 – Établissements industriels dans les Bouches-du-Rhône
Champ : établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône
Source : *répertoire SIRUS, 2013*
Note : La couleur est plus foncée lorsque le chiffre d'affaires est important

11.3.2 Modèle

Une entreprise peut être influencée dans son processus de production par la proximité géographique qu'elle entretient avec des entreprises voisines. Ces interactions sont regroupées sous le concept d' "externalités" qui peuvent être positives lorsque le voisinage a un impact favorable sur la production (complémentarités entre secteurs, intégration des chaînes de production, relation avec des fournisseurs, transport, partage de connaissance,..) ou négatives lorsqu'elles nuisent à la production (concurrence, pollution, embouteillages, etc.).

La production Y_i d'une entreprise i peut s'exprimer selon une loi de type Cobb-Douglas : $Y_i = AL_i^{\beta_L} K_i^{\beta_K}$, en fonction de son effectif moyen L_i, de son capital K_i et de la productivité générale des facteurs A. Les paramètres β_L et β_K représentent respectivement la part des revenus du travail et du capital dans la production [10]. Traditionnellement, le terme A désigne l'ensemble des mécanismes qui influencent la production (capital humain, progrès technologique, complémentarités...) sans pouvoir être directement mesurés. Il peut aussi être conçu comme représentant les externalités positives liées à la production et s'écrire : $A = \exp(\beta_0) \prod_{j \in v_i} Y_j^{\rho \omega_{ij}}$, où v_i désigne le voisinage de l'établissement i, Y_j le niveau de production d'une unité voisine de i. Le terme $\rho \omega_{ij}$ désigne l'élasticité de la production de l'entreprise i par rapport à celle de l'entreprise j : lorsqu'une entreprise j voisine de i augmente sa production de 1 %, la production de l'entreprise i augmente de $\rho \omega_{ij}$ %. Le paramètre ρ capte les complémentarités communes à toutes les unités tandis que ω_j capte les complémentarités spécifiques, résultant de l'impact de l'activité de j sur la production de i. En composant par la fonction logarithme, l'équation estimée peut se réécrire :

$$\log(Y_i) = \beta_0 + \rho \sum_{j \in v_i} \omega_{ij} \log(Y_j) + \beta_L \log(L_i) + \beta_K \log(K_i) + \varepsilon_i \tag{11.3}$$

On voit ainsi apparaître la forme d'une équation caractéristique d'un modèle spatial autorégressif (SAR), où la variable expliquée de l'observation i est régressée sur la somme pondérée des valeurs de cette variable chez les observations voisines de i. ρ peut alors être interprété comme le paramètre de corrélation spatiale. ω_{ij} représente la force de l'interaction entre les unités i et j : c'est le poids associé à ces unités dans le matrice de pondération spatiale.

10. Ces paramètres peuvent également être interprétés respectivement comme les élasticités de la production au travail et au capital.

11.3.3 Estimation

L'équation 11.3 est estimée sur 6 306 établissements géolocalisés dans les Bouches-du-Rhône, appartenant au secteur de l'industrie [11]. Ce secteur est particulièrement approprié à une estimation spatiale, car il ne fait pas directement intervenir la localisation géographique dans la production (contrairement au commerce, aux transports ou à l'agriculture), n'est pas trop concentré (comme les hautes technologies) et ne fait pas particulièrement intervenir des logiques de réseau autres que spatiales (comme en finance ou dans les communications).

La production Y_i d'un établissement est donnée par le chiffre d'affaires. Le total du bilan de l'entreprise, qui est une mesure de son patrimoine, sert de proxy pour le capital de l'établissement K_i. Ces deux variables, uniquement disponibles à l'échelle de l'entreprise, sont divisées par le nombre d'établissements au sein de l'entreprise. Enfin, l'effectif L_i est disponible au niveau de l'établissement dans SIRUS.

La figure 11.8 représente par des croix la localisation de ces établissements. L'intensité de la couleur verte de ces croix matérialise la taille de leur chiffre d'affaires : plus la couleur est foncée, plus le chiffre d'affaires est important. Des cliques d'établissements avec des forts chiffres d'affaires semblent se former, par exemple vers Aix-en-Provence ou autour de Fos-sur-Mer. De même que dans les simulations de la section 11.1, le voisinage des établissements est représenté par une matrice de poids fondée sur la distance. Selon notre définition, chaque établissement a en moyenne 109 voisins et 76 établissements n'ont pas de voisins [12].

β_0	β_L	β_K	ρ
0.422	0.535	0.769	0.051
(0.050)	(0.015)	(0.009)	(0.009)

TABLE 11.5 – Estimation du modèle SAR : ensemble des établissements
Champ : ensemble des établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône, dont le chiffre d'affaires et le total de bilan sont strictement positifs
Source : *répertoire SIRUS, 2015*
Note : Les paramètres estimés sont tous significatifs au seuil de 1 %.

La table 11.5 présente les résultats du modèle SAR estimé sur données exhaustives à l'échelle du département des Bouches-du-Rhône. Les parts des revenus du travail et du capital dans la production sont proches de celles généralement estimées (de l'ordre d'un demi à deux tiers pour la première, un tiers à deux tiers pour la seconde, le fort rendement marginal du capital pouvant ici s'expliquer par le choix du secteur industriel). Il existe bien une corrélation spatiale positive et significative : lorsque le chiffre d'affaires moyen des établissements voisins de i augmente de 1 %, le chiffre d'affaires de i augmente de 0,05 %.

11.3.4 Estimations spatiales sur des échantillons

Plans de sondage

De même que dans la section 11.1, nous répliquons l'estimation du modèle 11.3 sur un échantillon d'établissements. La sélection par sondage aléatoire simple sert de référence, mais n'est pas courante dans le cadre d'enquêtes auprès des entreprises. Les sondages stratifiés sont plus

11. Le secteur de l'industrie regroupe les établissements dont l'activité principale appartient aux divisions 10 à 33 de la NAF rév 2. 2008.

12. Ces unités sans voisins, aussi appelées "îles", ne participent donc pas à l'estimation du paramètre de corrélation spatiale ρ. Le choix du seuil résulte ainsi d'un arbitrage visant à minimiser à la fois le nombre de voisins et le nombre d'îles.

fréquemment employés dans le cadre d'études identifiant l'effet de l'effectif et du patrimoine sur le chiffre d'affaires. Ces méthodes de sondages ont été présentées en partie 11.1.2.

Dans le répertoire SIRUS, l'effectif est renseigné sur l'ensemble de la population. La stratification est effectuée selon cette variable d'effectif, sous l'hypothèse d'une corrélation entre effectif et chiffre d'affaires. La table 11.6 présente les strates ainsi constituées selon une allocation de Neyman, fondée sur la dispersion des chiffres d'affaires au sein de chacune des strates. La dispersion au sein de la strate 4 est bien supérieure à celle des autres strates, ce qui amène à considérer la strate 4 comme exhaustive, c'est-à-dire à toujours enquêter ces 67 établissements afin de limiter la variance d'estimation.

Numéro de strate	Nombre de salariés	Nombre d'établissements
1	0	3 628
2	1 à 9	2 742
3	10 à 99	770
4	100 et +	67

TABLE 11.6 – Définition des strates
Source : *répertoire SIRUS, 2013*
Champ : ensemble des établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône, dont le chiffre d'affaires et le total de bilan sont strictement positifs

Résultats

Dans cette partie, nous comparons les résultats obtenus avec un plan de sondage aléatoire simple et stratifié, en faisant varier la taille de l'échantillon : $n \in \{250, 500, 1000, 2000\}$.

	Échantillon aléatoire (sondage aléatoire simple)			Échantillon stratifié		
n	ρ	β_L	β_K	ρ	β_L	β_K
250	0.011	0.554^{***}	0.768^{***}	0.015^{*}	0.311^{***}	0.813^{***}
	(0.021)	(0.104)	(0.078)	(0.009)	(0.072)	(0.056)
500	0.017	0.545^{***}	0.773^{***}	0.020^{**}	0.371^{***}	0.796^{***}
	(0.016)	(0.073)	(0.052)	(0.008)	(0.053)	(0.041)
1000	0.024^{**}	0.542^{***}	0.774^{***}	0.024^{***}	0.410^{***}	0.793^{***}
	(0.012)	(0.051)	(0.039)	(0.007)	(0.039)	(0.029)
2000	0.034^{***}	0.541^{***}	0.770^{***}	0.036^{***}	0.457^{***}	0.790^{***}
	(0.010)	(0.031)	(0.023)	(0.007)	(0.028)	(0.022)

TABLE 11.7 – Modèle 11.3 estimé sur échantillon aléatoire (sondage aléatoire simple)
Note : régression non pondérée.
Champ : ensemble des établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône, dont le chiffre d'affaires et le total de bilan sont strictement positifs.
Source : *répertoire SIRUS, 2015*.

La table 11.7 présente les paramètres du modèle SAR estimés à partir de 1 000 tirages d'échantillon par sondage aléatoire simple (*à gauche*) et par sondage stratifié (*à droite*). Dans le cas d'un sondage aléatoire simple, de même que dans la section 11.1, les paramètres classiques de régression β_L et β_K, sont correctement estimés. En revanche, le paramètre de corrélation spatiale ρ n'est

significatif que pour un échantillon de taille supérieure à 1 000 et reste toujours inférieur à la valeur qu'il prend sur données exhaustives.

Le plan de sondage stratifié appliqué aux données non repondérées biaise les estimateurs classiques β_L et β_K lorsque la régression est non pondérée (DAVEZIES et al. 2009). En revanche, le biais sur le paramètre de corrélation spatiale ρ semble moindre. En effet, les grosses entreprises susceptibles d'avoir une influence spatiale importante sont toutes prises en compte dans l'échantillon du fait de ce plan de sondage stratifié.

Le choix de ne pas pondérer la régression est effectué par défaut. En économétrie classique, il est pertinent de pondérer les observations avant d'estimer un modèle économétrique lorsque la structure du plan de sondage est liée aux variables estimées. Cependant, la question de l'utilisation de poids de sondage dans le cadre d'un modèle de type SAR n'a pas été tranchée par la littérature actuelle [13]. En l'état actuel des choses, la régression non pondérée semble le choix le plus sûr et le plus simple à effectuer. Nous n'explorons pas plus avant cette question dans ce chapitre.

11.3.5 Estimation sur données agrégées

Tel qu'évoqué dans la partie 11.2, une approche couramment employée afin de contourner le problème de données manquantes est de passer à une échelle plus large par agrégation des données échantillonnées. Afin de s'abstraire des zonages administratifs, nous découpons le département des Bouches-du-Rhône selon une grille de taille $G \times G$ (figure 11.9).

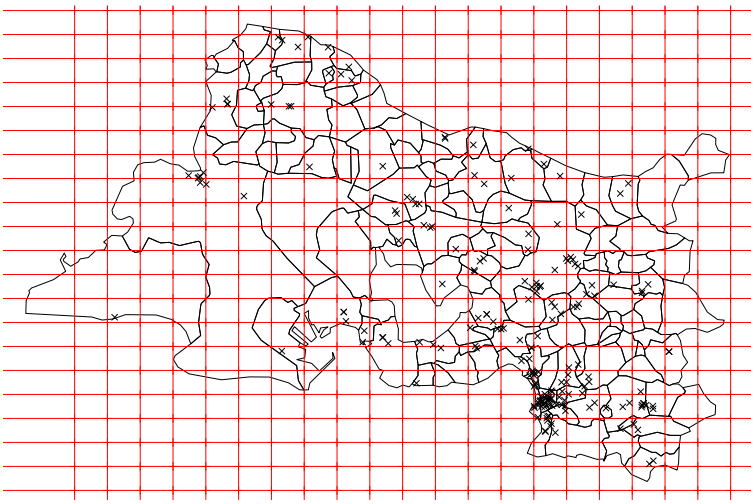


FIGURE 11.9 – Découpage des Bouches-du-Rhône selon une grille 20×20
Source : *répertoire SIRUS, 2013*
Champ : établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône

À partir de ce découpage, les observations d'un échantillon sont moyennées sur chaque cellule de la grille puis l'analyse spatiale est menée à l'échelle de la grille, les distances considérées étant définies entre centroïdes des cellules. Des valeurs nulles sont assignées aux variables et aux poids spatiaux des cellules sans observations, ce qui les exclut de fait de l'estimation sans distordre la taille de la matrice de pondération spatiale. La table 11.8 présente le paramètre ρ estimé pour différentes tailles d'échantillon et diverses tailles de grille.

13. Par exemple, il n'est pas clair s'il est nécessaire ou non de faire intervenir les poids de sondage dans le calcul de la matrice de pondération spatiale $\mathbf{W}$; cela pourrait également induire de l'endogénéité supplémentaire, liée à la structure de l'échantillon.

n \ G	20	30	50	60
100	0.007	0.009	0.014	0.015
	(0.018)	(0.022)	(0.022)	(0.024)
200	0.013	0.007	0.015	0.018
	(0.021)	(0.019)	(0.018)	(0.018)
500	0.024	0.023	0.012	0.013
	(0.031)	(0.023)	(0.014)	(0.013)
1000	0.031	0.057*	0.021*	0.014
	(0.026)	(0.040)	(0.015)	(0.012)

TABLE 11.8 – Paramètre ρ estimé sur données agrégées
Champ : ensemble des établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône, dont le chiffre d'affaires et le total de bilan sont strictement positifs
Source : *répertoire SIRUS, 2015*
Note : Les estimations sont réalisées sur les données d'un échantillon de taille n agrégées à l'échelle d'une grille de taille $G \times G$. Pour des raisons de lisibilité, nous ne représentons que les valeurs du paramètre ρ.

L'estimation de modèles spatiaux sur données agrégées semble permettre de contourner le problème des données manquantes dans un cadre très simple de données simulées de façon uniforme sur un territoire. Cependant, l'application de cette méthode à des données réelles n'est pas immédiate. En particulier, dans le cas présent, le paramètre d'autocorrélation spatiale est toujours sous-estimé et n'est jamais significatif. Cela pourrait être dû à la forte concentration de l'industrie des Bouches-du-Rhône, les distances intra-cellule n'étant par définition pas prises en compte par cette méthode. Les estimations spatiales sur données agrégées requièrent ainsi de s'assurer que le phénomène estimé n'est pas propre à une échelle géographique plus fine.

11.3.6 Imputation des données manquantes

Mise en œuvre

La seconde approche, évoquée en section 11.2.2, consiste à imputer les données manquantes, c'est-à-dire à attribuer des valeurs Y_i estimées aux établissements pour lesquels on n'en dispose pas. Nous considérons trois types d'imputation à l'échelle des établissements des Bouches-du-Rhône : *(i) l'imputation par le ratio*, faisant intervenir les variables L et K d'effectif et de capital comme variables explicatives du modèle, *(ii)* l'imputation par *hot deck statistique*, au sens où la distance est calculée en fonction des valeurs de L et de K, c'est à dire que les voisins d'un individu sont les établissements qui partagent des effectifs et des capitaux proches et *(iii)* l'imputation par *hot deck géographique*, où l'on associe à un individu ses voisins au sens géographique.

La mise en œuvre de ces techniques requiert, dans le premier cas, d'estimer un modèle linéaire (fonction `lm` de R) et dans les deux suivants, de définir les voisins (fonction `knn` du package *class* de R) puis de réaliser un tirage aléatoire parmi eux (fonction `sample` de R). Ces trois approches sont testées sur les données de l'industrie dans les Bouches-du-Rhône. 1 000 échantillons de taille n sont tirés selon un sondage aléatoire simple, puis le processus d'imputation assigne des valeurs de Y aux $N - n$ établissements non échantillonnés. Les résultats obtenus sont présentés dans la table 11.9. Pour rappel, les résultats sur la population entière sont en table 11.7.

	Ratio			Hot Deck Statistique			Hot Deck Géographique		
n	ρ	β_L	β_K	ρ	β_L	β_K	ρ	β_L	β_K
250	0.002	0.560***	0.768***	0.043***	0.664***	0.646***	0.419***	0.028	0.104***
	(0.002)	(0.112)	(0.080)	(0.009)	(0.083)	(0.059)	(0.046)	(0.034)	(0.023)
500	0.004	0.548***	0.774***	0.042***	0.613***	0.698***	0.412***	0.061*	0.149***
	(0.003)	(0.077)	(0.058)	(0.008)	(0.061)	(0.044)	(0.035)	(0.034)	(0.022)
1000	0.008**	0.546***	0.774***	0.040***	0.577***	0.734***	0.389***	0.116***	0.217***
	(0.003)	(0.051)	(0.037)	(0.007)	(0.040)	(0.028)	(0.035)	(0.035)	(0.023)
2000	0.017***	0.542***	0.773***	0.040***	0.562***	0.751***	0.333***	0.203***	0.338***
	(0.004)	(0.032)	(0.024)	(0.007)	(0.031)	(0.023)	(0.022)	(0.034)	(0.022)

TABLE 11.9 – Méthodes d'imputation
Source : *répertoire SIRUS, 2015*
Champ : ensemble des établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône, dont le chiffre d'affaires et le total de bilan sont strictement positifs

Résultats

Les résultats sont très variables selon la méthode utilisée. *L'imputation par le ratio* permet de bien conserver la structure linéaire entre chiffre d'affaires, effectif et capital, ce qui se traduit par des estimations sans biais et précises des coefficients β_L et β_K. En revanche, le ρ estimé est très faible, encore plus que dans le cas du sondage aléatoire simple exploité directement (voir table 11.7). En effet, l'imputation ne prend absolument pas en compte la structure spatiale, qui est effacée lors de l'estimation du modèle sur les données complétées. Il n'est donc pas pertinent d'essayer d'appliquer des modèles d'économétrie spatiale sur des données imputées avec cette méthode.

L'imputation par *hot deck statistique* semble plus prometteuse. Les estimateurs sont du bon ordre de grandeur par rapport aux valeurs obtenues sur la population et sont estimés avec précision. Une comparaison avec la table 11.5 révèle un biais lorsque $\hat{\rho}$, $\hat{\beta}_L$ et $\hat{\beta}_K$ sont estimés sur des échantillons de petite taille. Ainsi, l'imputation par hot deck biaise les estimateurs du modèle (CHARREAUX et al. 2016) mais permet de faire ressortir la structure des corrélations spatiales. En effet, le lien entre donneur et receveur semble conserver de façon implicite la structure des interactions spatiales. Il est également possible que la structure spatiale sous-jacente à Y existe aussi pour L et K et soit récupérée par imputation. Ainsi, l'emploi de cette méthode d'imputation à des fins d'analyse économétrique revient à un arbitrage entre biais et variance sur les paramètres β_L et β_K, classique en théorie des sondages. Cependant, dans le cas présent, la méthode présente en outre l'avantage de réduire considérablement le biais préexistant sur ρ. Ces résultats, testés uniquement sur ce jeu de données et sur un plan de sondage simple, sont à utiliser avec précaution. En tout état de cause, ce n'est pas sur la proximité spatiale entre donneur et receveur que repose l'efficacité de cette méthode, comme le montre le dernier exemple.

La méthode d'imputation par *hot deck géographique* conduit à des résultats aberrants. Se fondant directement sur la proximité spatiale entre donneur et receveur, elle entraîne une surestimation très forte de l'effet spatial (ρ très supérieur à la vraie valeur), au détriment de l'effet des autres variables du modèle (β_1 et β_2 très inférieurs aux vraies valeurs). En effet, selon cette méthode, des établissement spatialement proches auront le même chiffre d'affaires Y, ce qui crée *ex-nihilo* une très forte corrélation spatiale positive. L'utilisation de la dimension spatiale pour pallier le problème des données manquantes n'est pas immédiate. La table 11.12 en annexe 11.3.6 présente les résultats obtenus pour une imputation par hot deck géographique en se limitant aux établissements ayant des effectifs proches. Le paramètre ρ est moins surestimé mais les résultats restent très éloignés de

l'estimation sur données exhaustives. Il semblerait possible d'utiliser l'information géographique de façon parcimonieuse pour l'imputation, mais cela demanderait une analyse plus poussée du jeu de données et une bonne connaissance de sa structure spatiale.

Conclusion

Ce chapitre met en évidence les difficultés liées à l'application de modèles d'économétrie spatiale à des données échantillonnées. Deux écueils s'y opposent en particulier : *(i)* un "effet taille" par lequel l'estimation sur un échantillon distord la matrice de pondération spatiale, et *(ii)* un effet résultant de l'omission d'unités spatialement corrélées avec les unités observées. Ces deux effets tendent à sous-estimer l'ampleur de la corrélation spatiale. Néanmoins, ce biais est plus faible dans le cas d'un sondage par grappes et lorsque l'échantillon est plus important.

Les études empiriques résolvent généralement ce problème en ignorant les observations manquantes, en agrégeant les données à une échelle plus large ou en imputant les valeurs manquantes. La première solution n'est jamais souhaitable. Les deux autres sont loin d'être parfaites, la difficulté étant de reconstituer un ensemble d'information complexe à partir de peu d'observations. L'imputation par hot deck statistique est prometteuse, mais nous ne montrons pas sa validité dans un cas général.

Si cette problématique est vouée à se développer avec l'importance des réseaux sociaux et des données géolocalisées, l'estimation de modèles spatiaux sur des données échantillonnées reste rare. En l'état, il reste préférable de considérer des données exhaustives. Le présent chapitre met en garde contre les solutions trop expéditives, telles que l'agrégation des données à une échelle supérieure, les méthodes d'imputation simplistes ou la suppression des données manquantes. Lorsque qu'un échantillon relativement important est disponible, ou issu d'un sondage par grappes, une estimation spatiale pourrait alors être envisagée, en gardant à l'esprit que le paramètre de corrélation spatiale obtenu sera sans doute sous-estimé.

Annexe

Choix du modèle et de la matrice de voisinage

Les tables 11.10 et 11.11 présentent des résultats obtenus en termes d'estimation des paramètres de modèles SAR ou SEM *via* une méthode Monte Carlo selon différentes matrices de voisinage et différentes tailles d'échantillon.

	ρ			β		
$\mathscr{M}$ / n	2 voisins	5 voisins	Distance	2 voisins	5 voisins	Distance
50	0.020	−0.003	0.042	1.107***	1.050***	1.054***
	(0.110)	(0.172)	(0.043)	(0.115)	(0.095)	(0.125)
100	0.063	0.069	0.058*	1.112***	1.056***	1.054***
	(0.076)	(0.111)	(0.031)	(0.079)	(0.065)	(0.086)
150	0.097*	0.115	0.073**	1.107***	1.052***	1.049***
	(0.060)	(0.088)	(0.028)	(0.062)	(0.051)	(0.068)
250	0.150***	0.189**	0.101**	1.105***	1.050***	1.053***
	(0.047)	(0.065)	(0.026)	(0.049)	(0.040)	(0.052)

TABLE 11.10 – Modèle SAR - Estimation par Monte Carlo
Note : Les écarts types sont entre parenthèses.

	λ			β		
$\mathscr{M}$ / n	2 voisins	5 voisins	Distance	2 voisins	5 voisins	Distance
50	−0.025	−0.110	0.008	1.003***	1.003***	1.002***
	(0.167)	(0.287)	(0.193)	(0.115)	(0.113)	(0.112)
100	0.008	−0.027	0.024	1.003***	1.004***	1.003***
	(0.113)	(0.182)	(0.124)	(0.080)	(0.078)	(0.078)
150	0.023	0.002	0.034	0.998***	0.998***	0.998***
	(0.090)	(0.144)	(0.099)	(0.065)	(0.063)	(0.063)
250	0.047	0.042	0.052	1.000***	1.000***	1.000***
	(0.069)	(0.108)	(0.079)	(0.051)	(0.050)	(0.050)

TABLE 11.11 – Modèle SEM - Estimation par Monte Carlo
Note : Les écarts types sont entre parenthèses

Imputation par hot deck géographique stratifié

La table 11.12 donne les résultats obtenus pour une imputation par hot deck géographique en se limitant aux établissements ayant des effectifs proches, c'est à dire ceux de la même strate (définie dans la table 11.6) que l'établissement ayant une valeur manquante.

	Hot Deck Géographique Stratifié		
n	ρ	β_L	β_K
250	0.137	1.216	0.029
	(0.037)	(0.100)	(0.026)
500	0.148	1.192	0.071
	(0.031)	(0.077)	(0.025)
1000	0.156	1.121	0.149
	(0.026)	(0.061)	(0.025)
2000	0.148	0.542	0.279
	(0.019)	(0.048)	(0.025)

TABLE 11.12 – Une autre méthode d'imputation
Source : *répertoire SIRUS, 2015*
Champ : ensemble des établissements du secteur de l'industrie, dans le département des Bouches-du-Rhône, dont le chiffre d'affaires et le total de bilan sont strictement positifs

Références - Chapitre 11

ANSELIN, Luc (2002b). « Under the hood : Issues in the specification and interpretation of spatial regression models ». *Agricultural Economics* 27.3, p. 247–267.

ARBIA, Giuseppe, Giuseppe ESPA et Diego GIULIANI (2016). « Dirty spatial econometrics ». *The Annals of Regional Science* 56.1, p. 177–189.

ARDILLY, Pascal (1994). *Les techniques de sondage*.

BELOTTI, F., G. HUGHES et A. Piano MORTARI (2017a). « Spatial panel-data models using Stata ». *Stata Journal* 17.1, 139–180(42).

BOEHMKE, Frederick J., Emily U. SCHILLING et Jude C. HAYS (2015). *Missing data in spatial regression*. Rapp. tech. Society for Political Methodology Summer Conference.

BURT, Ronald S. (1987). « A Note on Missing Network Data in the General Social Survey ». *Social Networks* 9, p. 63–73.

CHARREAUX, C et al. (2016). « Econométrie et Données d'Enquête : les effets de l'imputation de la non-réponse partielle sur l'estimation des paramètres d'un modèle économétrique ».

CHOW, Gregory C. et An-Loh LIN (1976). « Best Linear Unbiased Estimation of Missing Observations in an Economic Time Series ». *Journal of the American Statistical Association* 71.355, p. 719–721.

CLIFF, A.D. et J.K. ORD (1972). *Spatial autocorrelation*. Pion, London.

COCHRAN, William G (2007). *Sampling techniques*. John Wiley & Sons.

DAVEZIES, L. et X. D'HAULTFOEUILLE (2009). *To Weight or not to Weight ? The Eternal Question of Econometricians facing Survey Data*. Documents de Travail de la DESE - Working Papers of the DESE g2009-06. INSEE, DESE.

DEMPSTER, A.P., N.M. LAIRD et D.B. RUBIN (1977). « Maximum likelihood from incomplete data via the EM algorithm ». *Journal of the royal statistical society* 39.1, p. 1–38.

EGGER, Peter et Michael PFAFFERMAYR (2006). « Spatial convergence ». *Papers in Regional Science* 85.2, p. 199–215.

ERTUR, Cem et Wilfried KOCH (2007). « Growth, technological interdependence and spatial externalities : theory and evidence ». *Journal of Applied Econometrics* 22.6, p. 1033–1062.

FERREIRO, Osvaldo (1987). « Methodologies for the estimation of missing observations in time series ». *Statistics and Probability Letters* 5.1, p. 65–69.

GILE, Krista J. et Mark S. HANDCOCK (2010). « Respondent-driven sampling : an assessment of current methodology ». *Sociological Methodology* 40.1, p. 285–327.

GOULARD, M., T. LAURENT et C. THOMAS AGNAN (2013). « About predictions in spatial autoregressive models : Optimal and almost optimal strategies ». *Toulouse School of Economics Working Paper* 13, p. 452.

HARVEY, A. C. et R. G. PIERSE (1984). « Estimating Missing Observations in Economic Time Series ». *Journal of the American Statistical Association* 79.385, p. 125–131.

HUISMAN, Mark (2014). *Imputation of missing network data*. Sous la dir. de Reda ALHAJJ et Jon ROKNE. T. 2. Springer, p. 707–715. ISBN : 978-1-4614-6169-2.

JONES, Richard H. (1980). « Maximum Likelihood Fitting of ARMA Models to Time Series with Missing Observations ». *Technometrics* 22.3, p. 389–395.

KELEJIAN, H.H. et I.R. PRUSHA (2010b). « Spatial models with spatially lagged dependent variables and incomplete data ». *Journal of geographical systems*.

KOSKINEN, Johan H., Garry L. ROBINS et Philippa E. PATTISON (2010). « Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation ». *Statistical Methodology* 7.3", p. 366–384.

KOSSINETS, Gueorgi (2006). « Effects of missing data in social networks ». *Social Networks* 28.3, p. 247–268.

LESAGE, James P., Manfred M. FISCHER et Thomas SCHERNGELL (2007). « Knowledge spillovers across Europe : Evidence from a Poisson spatial interaction model with spatial effects ». *Papers in Regional Science* 86.3, p. 393–421. ISSN : 1435-5957.

LESAGE, J.P. et R.K. PACE (2004). « Models for spatially dependent missing data ». *The journal of real estate finance and economics* 29.2, p. 233–254.

LITTLE, Roderick J. A. (1988). « Missing-Data Adjustments in Large Surveys ». *Journal of Business and Economic Statistics* 6.3, p. 287–296.

LITTLE, Roderick J. A. et Donald B. RUBIN (2002). *Statistical analysis with missing data*. 2nd. Wiley, Hoboken.

LIU, Xiaodong, Eleonora PATACCHINI et Edoardo RAINONE (2017). « Peer effects in bedtime decisions among adolescents : a social network model with sampled data ». *The Econometrics Journal*.

LÓPEZ-BAZO, Enrique, Esther VAYÁ et Manuel ARTÍS (2004). « Regional Externalities And Growth : Evidence From European Regions ». *Journal of Regional Science* 44.1, p. 43–73.

MARDIA, Kanti V. et al. (1998). « The Kriged Kalman filter ». *Test* 7.2, p. 217–282.

PINKSE, Joris et Margaret E. SLADE (2010). « The Future of Spatial Econometrics ». *Journal of Regional Science* 50.1, p. 103–117.

REVELLI, Federico et Per TOVMO (2007). « Revealed yardstick competition : Local government efficiency patterns in Norway ». *Journal of Urban Economics* 62.1, p. 121–134.

RUBIN, Donald B. (1976). « Inference and missing data ». *Biometrika* 63, p. 581–592.

STORK, Diana et William D. RICHARDS (1992). « Nonrespondents in Communication Network Studies ». *Group & Organization Management* 17.2, p. 193–209.

TILLÉ, Y. (2001). *Théorie des sondages : échantillonnage et estimation en populations finies : cours et exercices avec solutions :[2e cycle, écoles d'ingénieurs]*. Dunod.

WANG, W. et L.-F. LEE (2013a). « Estimation of spatial autoregressive models with randomly missing data in the dependent variable ». *The Econometrics Journal* 16.1, p. 73–102.

ZHOU, Jing et al. (2017). « Estimating Spatial Autocorrelation With Sampled Network Data ». *Journal of Business and Economic Statistics* 35.1, p. 130–138.

12. Petits domaines et corrélation spatiale

PASCAL ARDILLY
Insee

PAUL BOUCHE
Ensai - Sciences Po

WENCAN ZHU
Ensai

Résumé

Lorsque l'on veut diffuser les résultats d'une enquête sur des petites populations, en particulier s'il s'agit de zones géographiques restreintes, les faibles effectifs de l'échantillon recoupant ces populations peuvent conduire à des estimations trop imprécises. La théorie classique des sondages n'apporte pas de solution satisfaisante à ce problème et il faut donc faire appel à des techniques d'estimation spécifiques, fondées sur l'utilisation d'informations auxiliaires et sur des modèles plus ou moins complexes. Tous ces modèles ont en commun de formaliser la liaison entre la variable d'intérêt et les variables auxiliaires. La liaison linéaire est la forme la plus simple mais on trouve d'autres modèles, de nature non linéaire (modèle de Poisson, modèle logistique). La plupart des modèles isolent des effets locaux, propres aux domaines. On peut introduire des corrélations entre ces effets, d'autant plus fortes que les domaines sont géographiquement proches. Cette corrélation spatiale est alors de nature à améliorer la qualité des estimations localisées.

Ce chapitre est consacré à la présentation générale de la problématique appelée "estimation sur petits domaines" en portant une attention plus particulière à la prise en compte de la corrélation spatiale dans les modèles.

12.1 Mise en place du modèle

12.1.1 Contexte et objectifs

Les statisticiens d'enquête portent un intérêt particulier à l'estimation de paramètres θ inconnus, définis dans une population finie et généralement de grande taille. La plupart des paramètres sont des totaux ou des dérivés immédiats de totaux tels que des moyennes ou des proportions. Plus rarement, on trouve des fonctions non linéaires mais que l'on peut tout de même exprimer comme des fonctions de totaux (ratios, variances dans la population, coefficients de corrélation ou de régression). Selon le sujet de l'enquête, on peut aussi vouloir estimer des paramètres fortement non linéaires, comme des quantiles ou des indicateurs d'inégalités, lesquels ne s'écrivent pas comme des fonctions de totaux.

Les paramètres sont définis à partir d'une (ou plusieurs) variables(s) d'intérêt et se formalisent par des expressions impliquant en toute généralité l'ensemble des individus de la population U. On notera Y la variable d'intérêt, que l'on considérera par la suite comme unique. Les individus de U étant identifiés par l'indice i, si le paramètre θ est un total T, alors $T = \sum_{i \in U} Y_i$.

Lorsqu'on ne dispose pas de la valeur individuelle Y_i pour chaque individu i de U, on doit procéder à une estimation de T par sondage, c'est-à-dire à partir d'informations Y_i obtenues sur un échantillon répondant, noté s, inclus dans U. Le tirage de l'échantillon relève généralement d'un plan d'échantillonnage complexe, associant par exemple une stratification, des tirages à probabilités inégales et des tirages à plusieurs degrés. Certains paramètres d'intérêt ne sont pas définis sur la population U toute entière mais sur une sous-population, notée d. Une telle sous-population s'appelle un *domaine*, et on a alors affaire à une estimation sur domaine. Dans ce cas, le paramètre d'intérêt θ peut être le total sur le domaine, soit $T_d = \sum_{i \in d} Y_i$, qu'il faut estimer à partir des données collectées. La théorie classique des sondages attribue à chaque unité échantillonnée i un poids de sondage w_i, coefficient réel positif qui dépend de la méthode d'échantillonnage et de traitement de la non-réponse et qui vient "dilater" la valeur de la variable d'intérêt Y_i. Pour estimer un total T défini sur la population complète U, l'estimateur prend une forme linéaire $\hat{T} = \sum_{i \in s} w_i Y_i$. Pour estimer un total sur un domaine d, on se contente de restreindre la somme aux éléments de d sans toucher à leur pondération, soit $\hat{T}_d = \sum_{i \in s \cap d} w_i Y_i$. Si le paramètre θ est une moyenne sur d, notée désormais $\bar{Y}_d$ (ce qui inclut les proportions, qui sont des moyennes de variables booléennes), on estime la taille N_d du domaine par $\hat{N}_d = \sum_{i \in s \cap d} w_i$ (une taille est un total de valeurs individuelles constantes égales à 1) et on forme le ratio $\hat{\bar{Y}}_d = \hat{T}_d / \hat{N}_d$. Mais si on connaît N_d, on peut aussi utiliser l'estimation alternative $\hat{\bar{Y}}_d = \hat{T}_d / N_d$.

Dans tous les cas de figure, l'échantillonnage entraîne une erreur spécifique des estimateurs $\hat{T}_d$ et $\hat{\bar{Y}}_d$ que l'on résume au moyen de deux indicateurs appelés respectivement *biais* et *variance d'échantillonnage*. Considérons le cas de $\hat{\bar{Y}}_d$. Le biais désigne la différence entre l'espérance de $\hat{\bar{Y}}_d$, c'est-à-dire l'estimation attendue "en moyenne" compte tenu de l'aléa qui conduit à la constitution de s, et le paramètre $\bar{Y}_d$, tandis que la variance d'échantillonnage mesure la sensibilité de l'estimation $\hat{\bar{Y}}_d$ à l'échantillon répondant s. Un plan de sondage précis se traduit par un faible biais et une faible variance. Les estimateurs issus de la théorie des sondages et utilisés par les statisticiens d'enquête sont généralement sans biais ou ont un biais négligeable. La variance d'échantillonnage est une fonction décroissante de n_d, où n_d est la taille de l'échantillon répondant recoupant le domaine d, c'est-à-dire la taille de $s \cap d$. Lorsque n_d est suffisamment petit pour que les objectifs de qualité de l'estimation $\hat{\bar{Y}}_d$ ne soient pas atteints, on est face à un problème d'estimation sur *petit domaine*.

Pour traiter cette difficulté, lorsqu'il n'est plus possible d'augmenter la valeur de la taille de s, notée n, il faut créer un contexte théorique nouveau qui permette de rendre l'estimation finale du paramètre θ (total T_d ou moyenne $\bar{Y}_d$) moins sensible à l'échantillon répondant s (ou $s \cap d$, ce qui est équivalent). C'est une technique de *modélisation* qui va le permettre. Il s'agit de se placer dans un cadre hypothétique simplificateur de la réalité (c'est la définition générale d'un modèle). L'approche habituelle consiste à considérer que Y_i s'explique essentiellement par un ensemble de variables individuelles X_i connues pour chaque unité i de la population tout en impliquant quelques grandeurs δ *a priori* inconnues - les paramètres du modèle. Il suffira d'estimer ces grandeurs δ pour pouvoir en déduire n'importe quelle valeur inconnue Y_i (correspondant aux cas $i \notin s$), et donc *in fine* la valeur du paramètre θ.

La mise en œuvre de la modélisation passe fondamentalement par la disponibilité d'informations auxiliaires. Naturellement, on pense à des variables connues au niveau individuel sur l'intégralité de la population U. Supposons que l'information auxiliaire relative à l'individu i soit composée de p variables individuelles, notées $X_{i,1}, X_{i,2}, ..., X_{i,p}$, et partons du principe qu'il existe une liaison "suffisamment fiable" entre ces valeurs et la variable d'intérêt Y_i. Cette liaison est par construction considérée comme valable lorsqu'on l'applique à l'intégralité de la population U, sans autre connaissance que $X_{i,1}, X_{i,2}, ..., X_{i,p}$. Il est essentiel qu'elle reste valable si on se limite à l'échantillon répondant s, ce qui signifie que l'information qu'apporte l'appartenance à l'échantillon répondant ne doit pas amener le statisticien à modifier l'expression formelle de cette relation (plan de sondage dit "non informatif"). Dans un monde idéal où tout serait simple, il existerait une certaine fonction f telle que pour tout individu i de U on a $Y_i = f(X_{i,1}, X_{i,2}, ..., X_{i,p}; \delta)$ où δ est un paramètre vectoriel inconnu à ce stade, dit paramètre du modèle. Dans ce contexte parfait, la forme fonctionnelle de la fonction f est parfaitement connue mais elle est néanmoins paramétrée par δ. Si on parvient, grâce à l'information collectée à l'occasion de l'enquête, à estimer de façon satisfaisante le paramètre δ, on pourra prédire les valeurs Y_i de tous les individus i non échantillonnés (ou échantillonnés mais non-répondants) et donc prédire θ.

Le cadre traditionnel de la statistique d'enquête qu'est la théorie des sondages ne s'appuie sur aucune modélisation et considère que la variable d'intérêt Y n'est pas aléatoire (elle est donc déterministe). C'est la procédure de sélection de l'échantillon et le mécanisme de non-réponse qui introduisent un aléa et cet aléa permet de considérer tout estimateur, comme par exemple l'estimateur de la moyenne $\bar{Y}_d$, comme une variable aléatoire. Or, en présence d'une modélisation de Y, puisque la réalité n'est pas celle d'un monde idéal et simple, il ne serait pas raisonnable de supposer qu'il y a égalité entre la valeur Y_i et une quelconque valeur de type $f(X_{i,1}, X_{i,2}, ..., X_{i,p}; \delta)$, car la relation entre Y_i et $X_{i,1}, X_{i,2}, ..., X_{i,p}$ serait trop contrainte et donc non crédible. C'est pourquoi il faut considérer que la fonction f comprend une composante aléatoire U_i, dont la première caractéristique est d'être guidée par le hasard. On doit désormais abandonner l'environnement traditionnel de la théorie des sondages et considérer que *les variables Y sont des variables aléatoires*, telles que $Y_i = f(X_{i,1}, X_{i,2}, ..., X_{i,p}, U_i; \delta)$.

12.1.2 Le modèle linéaire individuel standard

Le formalisme du modèle utilise par conséquent des aléas explicites qui lui sont propres et qui n'ont aucune relation avec l'aléa d'échantillonnage. Dans certaines circonstances, on a coutume d'introduire une variable aléatoire individuelle U_i, qui est en moyenne nulle et que l'on relie ainsi à Y_i, pour tout i de U (équation 12.1).

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_p X_{i,p} + U_i \tag{12.1}$$

Les variables auxiliaires $X_{i,1}, X_{i,2}, ..., X_{i,p}$, parfaitement déterministes, sont dites "effets fixes". L'aléa du modèle portant sur les valeurs Y_i ne doit pas être confondu avec l'aléa de sondage qui

détermine la composition de l'échantillon s. C'est à ce stade que le contexte de l'estimation sur petits domaines apporte sa spécificité. La population U étant partitionnée en D domaines, on considère que si i appartient au domaine d, l'aléa U_i - nul *en moyenne* - est composé d'un effet (aléatoire) propre au domaine d, noté τ_d, et d'un résidu (aléatoire) individuel noté e_i. On a donc :

$$U_i = \tau_d + e_i. \tag{12.2}$$

Dans l'approche la plus simple, les deux composantes τ_d et e_i sont supposées indépendantes, les τ_d sont deux à deux indépendants, de même les e_i sont deux à deux indépendants. L'espérance et la variance associées à l'aléa du modèle seront notés respectivement ε et ν, si bien que les hypothèses les plus simples accompagnant ce modèle sont :

— pour les espérances : $\varepsilon(\tau_d) = 0$ et $\varepsilon(e_i) = 0$;
— pour les variances : $\nu(\tau_d) = \sigma_\tau^2$ et $\nu(e_i) = \sigma_e^2$.

Par ailleurs, toutes les covariances envisageables impliquant ces composantes élémentaires sont nulles. Ainsi, globalement $\varepsilon(U_i) = 0$ et $\nu(U_i) = \sigma_\tau^2 + \sigma_e^2$. Le formalisme de ce modèle permet de créer une corrélation entre les variables d'intérêt associées à des unités d'un même domaine puisque $\forall i \in U$, $\forall j \in U$, $j \neq i$: si $i \in d$ et $j \notin d$ alors $cov(Y_i, Y_j) = cov(U_i, U_j) = 0$ et si $i \in d$ et $j \in d$ alors $cov(Y_i, Y_j) = cov(U_i, U_j) = \sigma_\tau^2$. Ainsi, la matrice de variances-covariances du vecteur des Y_i, où i parcourt U, a la forme d'une matrice diagonale par blocs, chaque bloc étant associé à un domaine et pouvant être décrit par une diagonale comprenant partout $\sigma_\tau^2 + \sigma_e^2$ alors que tous les autres éléments du bloc prennent la valeur constante σ_τ^2.

Du fait des hypothèses portant sur les moments des aléas, un tel modèle ne peut s'appliquer, en toute rigueur, qu'à des variables Y quantitatives et continues - ce qui exclut en particulier toute variable d'intérêt de nature qualitative (et donc les paramètres définis comme des proportions). L'effet aléatoire τ_d est un effet local qui s'interprète comme étant la composante de la variable d'intérêt expliquée par l'appartenance au domaine au-delà de l'information contenue dans les variables individuelles $X_{i,1}, X_{i,2}, ..., X_{i,p}$. Souvent, les domaines sont des zones géographiques et τ_d prétend traduire la part d'explication purement due à la localisation géographique de l'unité. Apprécier la vraie part explicative de la localisation sur telle ou telle zone, et même définir ce qu'est un effet géographique, constitue une question un peu philosophique. En effet, parce que c'est une explication facile et bien pratique, on peut toujours considérer comme effet géographique un effet résiduel significatif qui serait dû à une insuffisante prise en compte des variables auxiliaires individuelles réellement explicatives. Autrement dit, s'il y a des éléments géographiques qui expliquent Y, l'idéal consiste à les traduire d'une façon ou d'une autre dans le vecteur d'effets fixes $X_{i,1}, X_{i,2}, ..., X_{i,p}$. Il faut donc concevoir *a priori* l'effet local τ_d comme un effet "parasite" et chercher à en diminuer au maximum l'importance : plus le paramètre σ_τ^2 sera petit, c'est-à-dire plus les valeurs τ_d seront numériquement faibles, plus le caractère explicatif reposera sur les effets fixes $X_{i,1}, X_{i,2}, ..., X_{i,p}$, et donc meilleur sera le modèle. La structure de covariance ayant une certaine complexité, on dit que le modèle appartient à la famille des modèles linéaires généraux.

Avec un tel modèle, l'espérance de la variable aléatoire Y_i est une fonction linéaire des paramètres β. La composante explicative de Y_i la plus importante est constituée d'effets non aléatoires $X_{i,j}$ (les effets fixes) mais la composante résiduelle τ_d attribuée exclusivement au domaine est en revanche de nature aléatoire (l'effet aléatoire). Pour ces raisons, on parle de *modèle linéaire mixte*.

Si on reprend les notations de la partie 12.1.1, on vérifie bien que $Y_i = f(X_{i,1}, X_{i,2}, ..., X_{i,p}, U_i; \delta)$, où le paramètre vectoriel δ rassemble toutes les grandeurs inconnues apparaissant dans le modèle, à savoir $\delta = (\beta_0, \beta_1, \beta_2, ..., \beta_p, \sigma_\tau^2, \sigma_e^2)$. Il a une dimension $p+3$, distinguant $p+1$ paramètres réels associés aux effets fixes explicatifs et deux paramètres réels associés à la structure de variances-covariances attachée au modèle.

12.1.3 Le modèle linéaire individuel avec corrélation spatiale

Le modèle linéaire mixte standard formule l'hypothèse d'une corrélation nulle entre les aléas U_i associés à des individus appartenant à deux domaines distincts. Cette situation n'est pas nécessairement crédible parce que les limites des zonages géographiques constituant les domaines n'ont aucune raison de constituer une barrière stoppant brutalement toute propagation des phénomènes mesurés. En général, il y a une forme de continuité spatiale naturelle des comportements des individus localisés et deux individus géographiquement proches sur le terrain ont plus de chance d'afficher des valeurs de Y voisines que deux individus éloignés. De ce point de vue, une relation entre les effets géographiques caractérisant des domaines proches apparaît assez naturelle.

Sur le plan technique, on peut chercher à traduire cette situation en introduisant une corrélation qui ne tienne plus compte que de la distance entre les domaines. La forme analytique de la corrélation est libre, pourvu qu'elle diminue quand la distance augmente. Dans cet esprit, on peut s'appuyer sur un modèle qui conserve exactement les formalisations 12.1 et 12.2 mais qui assure $\forall i \in d, \forall j \in d'$, si $i \neq j$:

$$cov(Y_i, Y_j) = cov(U_i, U_j) = cov(\tau_d, \tau_{d'}) = \sigma_\tau^2 exp\left(-\frac{1}{\rho} dist(d, d')\right) \tag{12.3}$$

où $dist(d, d')$ est une distance définie entre les domaines d et d'. On peut prendre par exemple la distance euclidienne habituelle calculée à partir des coordonnées des centroïdes des deux domaines concernés. Le coefficient ρ est un paramètre d'échelle qui offre l'opportunité d'un meilleur ajustement du modèle : plus la distance sera influente sur la covariance, plus ρ sera proche de zéro. Dans le cas particulier où $d = d'$, et lorsque $i \neq j$, alors $cov(Y_i, Y_j) = cov(\tau_d, \tau_d) = \sigma_\tau^2 exp(0) = \sigma_\tau^2$. Si $i = j$, s'ajoute la variance de l'effet individuel, soit $cov(Y_i, Y_j) = cov(U_i, U_j) = \sigma_\tau^2 + \sigma_e^2$. Cette fois, la matrice de variances-covariances est une matrice pleine, sans zéros. On peut néanmoins considérer, à titre de variante intéressante, que la distance devient infinie lorsqu'elle a dépassé un certain seuil. Cela permet de réintroduire de nombreux zéros dans la matrice, facilitant ainsi les traitements numériques ultérieurs (en particulier en épargnant de la mémoire vive). Dans ces circonstances, les paramètres du modèle sont un peu plus nombreux puisqu'il faut tenir compte du nouveau paramètre ρ, si bien que $\delta = (\beta_0, \beta_1, \beta_2, ..., \beta_p, \rho, \sigma_\tau^2, \sigma_e^2)$.

Une autre approche consiste à introduire une relation simple entre les effets locaux τ_d des différents domaines, en faisant en sorte que cette relation soit d'autant plus forte que les domaines sont plus proches. Ainsi, on peut concevoir que l'effet local associé à un domaine donné soit "presque" une combinaison linéaire des effets locaux des domaines qui l'entourent, avec une intensité de liaison qui diminue au fur et à mesure qu'on s'éloigne du domaine donné. L'intensité de la liaison entre les effets τ_d est traduite par deux éléments. D'une part un système de coefficients $\alpha_{d,d\prime}$ [1] qui règlent l'influence relative que peuvent avoir les différents domaines distingués $d\prime$ sur un domaine donné d, d'autre part un paramètre ρ compris entre -1 et 1 qui règle la valeur absolue de l'intensité de liaison. On impose pour tout i : $\sum_{d\prime=1, d\prime\neq d}^{D} \alpha_{d,d\prime} = 1$. La relation postulée entre les effets aléatoires est $\tau_d \approx \rho \sum_{d\prime=1, d\prime\neq d}^{D} \alpha_{d,d\prime} \tau_{d\prime}$. En écriture matricielle, cela devient :

$$\begin{pmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_D \end{pmatrix} = \rho . \begin{pmatrix} 0 & \alpha_{1,2} & \dots & \alpha_{1,D} \\ \alpha_{2,1} & 0 & \dots & \alpha_{2,D} \\ \vdots & \ddots & 0 & \vdots \\ \alpha_{D,1} & \dots & \dots & 0 \end{pmatrix} \begin{pmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_D \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_D \end{pmatrix} \tag{12.4}$$

1. Les paramètres $\alpha_{d,d\prime}$ correspondent aux poids $w_{d,d\prime}$ de la matrice de poids W utilisée dans les chapitres précédents. Dans ce chapitre, w désigne les poids de sondage.

en introduisant un vecteur d'aléas u_d qui suit une loi de Gauss centrée et de variance $\sigma_u^2 I_D$. On désigne ce modèle sous le nom de modèle SAR (*Simultaneous AutoRegressive model*).

L'arbitrage entre cette méthode et la précédente n'a rien d'évident *a priori*, c'est pourquoi le seul conseil à prodiguer à ce stade consiste à tester les deux méthodes puis à utiliser les outils d'appréciation de la qualité dont on dispose, en particulier ceux mentionnés dans la partie 12.3.

L'introduction d'une corrélation spatiale dans le modèle linéaire mixte de base ne change rien aux conditions restrictives d'usage : un tel modèle ne peut être utilisé que pour l'estimation de paramètres θ construits à partir d'une variable d'intérêt quantitative et continue. En outre, il perd une grande partie de son intérêt si les domaines sont géographiquement de grande taille car la distance considérée se mesure entre les centroïdes des domaines.

En pratique, pour limiter le nombre de coefficients non nuls dans la matrice de variances-covariances des effets locaux (et ainsi accélérer les calculs et/ou éviter des problèmes de mémoire insuffisante), on neutralise complètement l'influence $\alpha_{d,d\prime}$ des domaines $d\prime$ situés au-delà d'une certaine distance de d, ou même éventuellement qui ne sont pas dans un voisinage immédiat du domaine de référence d. Néanmoins, il est difficile d'éviter les problèmes posés par les "effets de bord" qui surviennent lorsqu'un domaine se trouve en périphérie d'un territoire plus vaste, parce qu'on ne peut pas prendre en compte tous ses voisins. C'est par exemple presque systématiquement le cas pour les territoires frontière des États.

12.1.4 Traitement des variables qualitatives par un modèle individuel linéaire mixte généralisé

Le modèle logistique

Les paramètres de dénombrement d'une sous-population quelconque s'appuient sur des variables individuelles de nature qualitative. Supposons que l'on cherche à estimer le nombre total d'individus θ vérifiant une propriété donnée Γ - comme par exemple "être une femme" ou "être un agriculteur de moins de 50 ans". Si on définit la variable individuelle $Y_i = 1$ lorsque i vérifie Γ et $Y_i = 0$ dans le cas contraire, il est facile de vérifier que $\theta = \sum_{i \in U} Y_i$. La variable aléatoire Y ainsi définie est une variable dite "indicatrice" qui quantifie une information individuelle initialement qualitative. En divisant θ par la taille de U, on obtient la proportion d'individus de la population qui vérifient la propriété Γ. Malheureusement, le modèle 12.1 n'est pas du tout adapté à ce type de variable. On va contourner la difficulté en optant pour une modélisation parfaitement compatible avec les variables indicatrices : la loi de Bernoulli traduira la distribution des Y_i. Il s'agit d'une distribution qui charge la valeur 1 avec une probabilité P_i et la valeur 0 avec une probabilité $1 - P_i$. On peut donc considérer que pour tout individu i de la population globale U, la variable Y_i est une variable aléatoire qui suit une loi de Bernoulli $\mathscr{B}(1, P_i)$. Le cœur de la modélisation suit : on va relier le paramètre P_i aux caractéristiques individuelles de i résumées par les variables auxiliaires $X_{i,1}, X_{i,2}, ..., X_{i,p}$ et on va introduire un effet aléatoire local τ_d. La forme fonctionnelle qui relie P_i aux $X_{i,1}, X_{i,2}, ..., X_{i,p}$ et à τ_d doit être compatible avec la contrainte $P_i \in [0,1]$. Différentes options existent, mais la plus commune consiste à poser, pour tout i dans d :

$$log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_p X_{i,p} + \tau_d. \tag{12.5}$$

On parle de *modèle logistique*. L'espérance de la variable aléatoire Y_i est P_i, qui n'est manifestement pas une fonction linéaire des paramètres β (contrairement au cas de 12.1). Pour cette raison, on dit que le modèle représenté par l'équation 12.5 est un *modèle linéaire mixte généralisé*. La classe des modèles de type 12.5 distingue les modèles où les effets locaux τ_d sont deux à deux indépendants, comme dans 12.2, et les modèles avec corrélation spatiale, comme dans 12.3 ou 12.4.

Le modèle de Poisson

Il arrive que l'information qualitative se présente de manière agrégée lorsqu'on traite les unités statistiques. Si on reprend l'exemple précédent, dans le cas où les unités sont des ménages et non plus des individus physiques, on dispose pour chaque ménage i du nombre total d'individus Y_i vérifiant la propriété Γ (le nombre de femmes dans le ménage, ou le nombre d'agriculteurs de moins de 50 ans dans le ménage). Cette variable n'est plus une variable indicatrice mais une variable qui peut prendre n'importe quelle valeur dans $\mathbb{N}$, ensemble des entiers naturels (en théorie, puisqu'en pratique elle est toujours bornée supérieurement). Dans ces conditions, la loi de Poisson est une loi naturelle assez simple que l'on peut associer à Y_i. Elle possède un unique paramètre λ_i réel (strictement positif), que l'on va faire dépendre de l'unité i au travers de caractéristiques individuelles $X_{i,1}, X_{i,2}, ..., X_{i,p}$ et d'un effet aléatoire local τ_d. Le paramètre λ_i est souvent transformé par une fonction simple avant d'être relié aux facteurs explicatifs. En pratique, on utilise essentiellement la fonction logarithme, ce qui fait que le modèle complet - linéaire mixte généralisé - se formalise ainsi :

$$\begin{aligned} Y_i &\sim Poisson(\lambda_i) \\ log(\lambda_i) &= \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_p X_{i,p} + \tau_d. \end{aligned} \tag{12.6}$$

Une fois encore, les effets aléatoires locaux τ_d peuvent être considérés comme deux à deux indépendants, comme dans 12.2, ou corrélés spatialement, comme dans 12.3 ou 12.4.

12.1.5 Extension aux modèles définis au niveau du domaine

Le modèle de Fay et Herriot

En tenant compte de l'échantillonnage, on peut produire des estimateurs de n'importe quel paramètre, en particulier les totaux T_d (ou les moyennes $\bar{Y}_d$) définis au niveau du domaine d. Ces estimateurs sont construits avec les poids de sondage individuels w_i (eux-mêmes fonction de la méthode d'échantillonnage utilisée). Ils n'utilisent que l'information relative au domaine d, c'est pourquoi on les appelle *estimateurs directs*. Il est possible de construire une modélisation qui s'appuie sur ces estimateurs, notés $\hat{T}_d$ pour les totaux et $\hat{\bar{Y}}_d$ pour les moyennes. L'unité statistique modélisée n'est plus alors l'individu mais le domaine. L'objectif est de relier l'information disponible $\hat{T}_d$ ou $\hat{\bar{Y}}_d$ à un ensemble de variables explicatives, ces dernières étant adaptées au niveau traité : il faut naturellement qu'elles caractérisent les domaines et non plus les individus. Les effets locaux τ_d conservent leur nature et leur interprétation, exactement comme dans l'équation 12.2.

Un célèbre modèle est le modèle dit de *Fay et Herriot*, qui appartient à la famille des modèles linéaires mixtes. Si on note $X_{d,1}, X_{d,2}, ..., X_{d,p}$ les variables explicatives retenues au niveau domaine, la version la plus élémentaire de la modélisation s'écrit :

$$\bar{Y}_d = \beta_0 + \beta_1 X_{d,1} + \beta_2 X_{d,2} + ... + \beta_p X_{d,p} + \tau_d. \tag{12.7}$$

La variable expliquée est ici la vraie moyenne dans le domaine d. Puisque cette valeur est inconnue, il faut ajouter une étape pour lui substituer une estimation. À ce stade, l'estimation $\hat{\bar{Y}}_d$ issue de l'enquête n'est certes pas de bonne qualité puisque l'échantillon $s \cap d$ est de petite taille, néanmoins elle existe et peut être reliée à la vraie valeur en introduisant un terme d'erreur err_d selon :

$$\hat{\bar{Y}}_d = \bar{Y}_d + err_d. \tag{12.8}$$

La variable err_d est l'erreur d'échantillonnage. Cette dernière équation n'a rien à voir avec un modèle, il s'agit simplement de la définition de l'erreur d'échantillonnage. Généralement

l'estimateur $\hat{\bar{Y}}_d$ est pondéré de façon à être sans biais ou de biais négligeable (s'il y a eu redressement par exemple, et si toutefois on considère que la non-réponse a été correctement traitée) si bien que l'erreur d'échantillonnage a une espérance nulle lorsqu'on prend en compte l'aléa de sondage, soit $\mathbb{E}(err_d) = 0$. La variance de l'erreur dépend de l'échantillonnage mais on sait qu'elle varie comme l'inverse de n_d. On notera désormais ψ_d cette variance. La combinaison des deux équations précédentes conduit à la formule opérationnelle :

$$\hat{\bar{Y}}_d = \beta_0 + \beta_1 X_{d,1} + \beta_2 X_{d,2} + ... + \beta_p X_{d,p} + \tau_d + err_d. \tag{12.9}$$

À l'image de ce que l'on a vu avec les modèles individuels, on peut formuler une hypothèse d'indépendance entre les effets locaux τ_d, ou au contraire postuler une corrélation spatiale, structurée comme dans les équations 12.3 ou 12.4.

Les hypothèses sur l'espérance et la variance des effets τ_d n'étant en toute rigueur compatibles qu'avec des vraies moyennes $\bar{Y}_d$ qui ont des distributions continues, autant dire que la variable d'intérêt individuelle Y_i collectée au niveau des individus devrait être quantitative et continue. Cela étant, si la variable d'intérêt Y_i est par nature qualitative et si le domaine d a une taille N_d suffisamment grande, on peut considérer - avec un peu d'audace parfois ! - que la vraie moyenne $\bar{Y}_d$ peut prendre *a priori* un nombre suffisamment grand de valeurs pour que cet ensemble puisse être considéré comme continu, c'est-à-dire sans "trou". C'est bien la taille N_d qui est le paramètre essentiel. Considérons par exemple Y_i la variable indicatrice caractérisant la modalité "femme". La moyenne $\bar{Y}_d$ est alors la proportion de femmes dans la population du domaine. Si $N_d = 10$, cette moyenne peut prendre les valeurs $k/10$, où k est un entier compris entre 0 et 10, ce qui est très loin d'occasionner une situation "continue". Si $N_d = 10\,000$, la moyenne peut prendre les valeurs $k/10\,000$, où k est un entier compris entre 0 et 10 000, ce qui rend beaucoup plus plausible l'hypothèse de continuité. C'est pourquoi on peut conclure que la modélisation 12.9 est acceptable pour l'estimation de proportions (variables d'intérêt qualitatives) dès lors que les domaines d ne sont pas trop petits.

Le modèle de Poisson

Bien que le modèle de Fay et Herriot s'accommode bien des variables qualitatives, c'est-à-dire des paramètres qui se définissent comme des proportions par domaine d'individus vérifiant une propriété Γ (assimilable à une sous-population Γ) ou comme les effectifs par domaine de ces mêmes individus, on peut lui préférer dans certaines circonstances un modèle plus spécifiquement adapté aux dénombrements. Notons $N_{\Gamma,d}$ le nombre total d'individus du domaine d appartenant à la sous-population Γ. L'échantillon permet de former l'estimateur sans biais (ou presque) $\hat{N}_{\Gamma,d} = \sum_{i \in s \cap d \cap \Gamma} w_i$. Cet estimateur n'utilise que l'information liée au domaine, c'est donc un estimateur direct, et il est de qualité médiocre puisque l'échantillon $s \cap d$ est de petite taille. Néanmoins, il s'agit d'une variable aléatoire calculable dont on peut modéliser la distribution par une loi de Poisson. Cette loi, dépendant d'un unique paramètre réel λ_d fonction du domaine, est particulièrement adaptée aux dénombrements. On peut montrer que λ_d est l'espérance mathématique de $\hat{N}_{\Gamma,d}$: il doit donc être numériquement assez proche de cette estimation. Il s'agit bien à ce stade d'une première hypothèse et non d'une propriété qui découlerait de la théorie des sondages. Néanmoins, le risque pris reste modeste parce que le comportement asymptotique des estimateurs directs est proche d'une loi de Gauss, dont la loi de Poisson est elle-même proche si son paramètre est suffisamment grand.

Le cœur de la modélisation relève de la suite : on considère généralement que le logarithme du paramètre λ_d s'écrit ainsi :

$$log(\lambda_d) = \beta_0 + \beta_1 X_{d,1} + \beta_2 X_{d,2} + ... + \beta_p X_{d,p} + \tau_d \tag{12.10}$$

en reprenant les notations des parties précédentes. La variable aléatoire τ_d conserve la même interprétation : il s'agit de distinguer l'effet de la localisation des unités statistiques au-delà de ce que les effets fixes $X_{d,1}, X_{d,2}, ..., X_{d,p}$ sont capables de traduire. Les hypothèses portant sur les corrélations entre les effets locaux τ_d sont identiques à celles des modèles déjà présentés : ou bien on considère que ces effets sont deux à deux indépendants, ce qui est plus simple mais peut-être parfois en décalage avec la réalité du terrain, ou bien on introduit des corrélations spatiales, en reprenant par exemple les formulations 12.3 ou 12.4. Dans les deux cas, le modèle est un modèle linéaire mixte généralisé.

12.2 Formation de l'estimateur "petits domaines"

Définir le modèle sur lequel on va s'appuyer ne constitue qu'une première étape du processus. À ce stade, on ne perçoit encore qu'assez qualitativement l'intérêt du modèle, qui consiste à réduire la dimension du problème en simplifiant considérablement la réalité. Il est en effet beaucoup plus facile de procéder à des estimations dans un univers où toute l'information d'intérêt est supposée s'expliquer par quelques variables bien connues et par quelques paramètres plutôt qu'à évoluer dans un système non cadré qui de fait dépendrait d'une infinité de composantes non maîtrisées... comme le suppose au demeurant la théorie classique des sondages !

L'étape suivante est celle du choix de la stratégie d'estimation - on devrait d'ailleurs désormais parler de prédiction puisque le paramètre d'intérêt est devenu une variable aléatoire à la suite de la modélisation.

12.2.1 Stratégie d'estimation BLUP : cas du modèle individuel standard

Dans cette partie, on considère uniquement des modèles linéaires. Dans ce cadre, plusieurs stratégies d'estimation/prédiction du paramètre d'intérêt peuvent être mises en œuvre mais nous présentons maintenant celle qui est probablement la plus commune, la stratégie *Best Linear Unbiased Predictor* (BLUP). Considérons le cas où le paramètre est la moyenne $\bar{Y}_d$. Son prédicteur est en toute généralité une fonction des données collectées, c'est-à-dire des Y_i où i décrit l'échantillon global répondant s. Le statisticien cherche avant tout un prédicteur qui soit linéaire, du type $\sum_{i\in s} a_i Y_i$ où les a_i sont des coefficients réels, et sans biais, c'est-à-dire que son espérance soit égale à celle de $\bar{Y}_d$. Enfin, il cherche à minimiser l'erreur quadratique moyenne (*Mean Square Error* ou MSE) qui est l'espérance du carré de l'écart entre le prédicteur et la valeur $\bar{Y}_d$ qu'il doit prédire. La solution de ce problème mathématique est l'estimateur (ou prédicteur) BLUP, dit aussi dans la littérature estimateur de Henderson. On le notera $\tilde{\bar{Y}}_d^H$.

Dans le cas spécifique du modèle linéaire mixte individuel standard (voir partie 12.1.2), lorsque la fraction de sondage est négligeable, on vérifie que l'estimateur BLUP s'écrit :

$$\tilde{\bar{Y}}_d^H = \gamma_d[\bar{y}_d + (\bar{X}_d - \bar{x}_d)^T \tilde{\beta}] + (1-\gamma_d)\bar{X}_d^T \tilde{\beta} \tag{12.11}$$

Tous les vecteurs sont des vecteurs colonnes, le vecteur transposé étant repéré par l'exposant T. En notant D le nombre total de domaines d'intérêt, en notant $X_i = (X_{i,1}, X_{i,2}, ..., X_{i,p})^T$ le vecteur des variables auxiliaires, $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_p)^T$ le vecteur des paramètres de modèle associés à ces variables, $\bar{x}_d = \frac{1}{n_d}\sum_{i\in s\cap d} X_i$, $\bar{y}_d = \frac{1}{n_d}\sum_{i\in s\cap d} Y_i$ et $\bar{X}_d = \frac{1}{N_d}\sum_{i=1}^{N_d} X_i$, on a :

$$\gamma_d = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \frac{\sigma_e^2}{n_d}} \tag{12.12}$$

$$\tilde{\beta} = \left(\sum_{d=1}^{D} \left(\sum_{i \in s \cap d} X_i X_i^T - \gamma_d n_d \bar{x}_d \bar{x}_d^T \right) \right)^{-1} . \left(\sum_{d=1}^{D} \left(\sum_{i \in s \cap d} X_i Y_i - \gamma_d n_d \bar{x}_d \bar{y}_d \right) \right). \quad (12.13)$$

Le vecteur de coefficients $\tilde{\beta}$ n'a pas ici une expression familière, mais on peut vérifier qu'il s'agit de l'estimateur classique et bien connu dit des "moindres carrés généralisés", rencontré fréquemment dans la théorie des modèles de régression linéaire. Il estime de manière optimale le vecteur de paramètres inconnus β du modèle.

Il est fondamental de noter qu'on a besoin de connaître les vraies moyennes par domaine $\bar{X}_d$. En pratique, cela signifie que les variables individuelles X_i sont disponibles dans un certain fichier exhaustif couvrant le champ de l'enquête (ce qui ne signifie pas que ces valeurs individuelles soient accessibles au statisticien en charge de l'estimation, qui ne dispose peut-être que des $\bar{X}_d$). Toutefois, il est possible que ce fichier ne soit pas la base de sondage et que l'information X_i mobilisée pour le calcul de $\tilde{\beta}$ provienne du fichier de collecte de l'enquête, exactement au même titre que Y_i. Dans ce cas, qui est courant en pratique, il convient de s'assurer que la variable X relève bien des mêmes concepts dans les deux sources (fichier exhaustif et fichier de collecte). Par exemple, calculer $\tilde{\beta}$ à partir d'une enquête Emploi où X représente le statut d'activité collecté dans l'enquête et former $\tilde{\bar{Y}}_d^H$ en utilisant des $\bar{X}_d$ qui représentent les statuts d'activité déclarés dans le recensement, s'avérerait très périlleux.

Formellement, l'estimateur de Henderson est constitué de deux éléments qui sont combinés grâce au coefficient réel γ_d. Le premier élément - situé entre les crochets de l'équation 12.11 - est un estimateur de circonstance dont l'interprétation est un peu compliquée mais qui a les mêmes performances statistiques que $\bar{y}_d$, estimateur construit à partir du sous-échantillon $s \cap d$: il a une variance d'échantillonnage fonction décroissante de n_d, donc *a priori* grande. Parce que cette caractéristique est associée aux estimateurs directs, mais qu'en même temps la présence du coefficient $\tilde{\beta}$, formé à partir de l'échantillon complet, ne permet pas de le qualifier rigoureusement d'estimateur direct, on parlera d'*estimateur pseudo direct*. Le second élément est un estimateur construit en multipliant le coefficient de régression $\tilde{\beta}$ par la vraie moyenne de la variable auxiliaire $\bar{X}_d$, ce qui intuitivement devrait donner une valeur proche de la vraie moyenne de la variable d'intérêt si le modèle est pertinent. Cet estimateur $\bar{X}_d^T \tilde{\beta}$ s'appelle *estimateur synthétique*. Ses propriétés statistiques sont totalement dépendantes de celles de $\tilde{\beta}$ puisque la moyenne $\bar{X}_d$ n'a aucun caractère aléatoire. Or on constate *de visu* que $\tilde{\beta}$ est constitué de termes impliquant l'intégralité de l'échantillon répondant s et non pas seulement la partie $s \cap d$. Cela lui confère par nature une grande stabilité, autrement dit une faible dépendance à l'échantillon répondant s. Si on ne considère que l'aléa de sondage, on peut donc dire que la composante synthétique offre une faible variance d'échantillonnage. La contrepartie de cette stabilité est l'existence d'un biais d'échantillonnage, qui peut être numériquement fort si le modèle est inadapté.

Le coefficient γ_d, qui est toujours compris entre 0 et 1, est un coefficient remarquable parce qu'il pondère de manière optimale (on rappelle qu'il minimise la MSE) les deux composantes distinguées, lesquelles ont des comportements tout à fait opposés en matière à la fois de biais et de variance d'échantillonnage. En cela, on dit que $\tilde{\bar{Y}}_d^H$ est un *estimateur composite* (ou *mixte*). La stratégie BLUP conduit donc à une expression de γ_d qui donne priorité à celle des deux composantes qui est la plus efficace. Prenons le cas où σ_τ^2 est petit, ce qui correspond à des effets locaux τ_d petits, autrement dit à un modèle performant puisqu'il fait porter le véritable caractère explicatif sur les variables auxiliaires maîtrisées X_i et non sur le terme résiduel "attrape-tout" τ_d. Dans de telles circonstances, on a tendance à faire confiance au modèle et à construire l'estimateur final en s'appuyant au maximum sur le modèle, c'est-à-dire sur l'estimateur synthétique. C'est effectivement ce qu'il advient puisque γ_d est petit. Prenons maintenant le cas où la taille d'échantillon répondant

n_d est grande. Un tel contexte donne confiance dans l'estimateur pseudo direct, qui n'utilise pas (ou très peu) le modèle, et donc par construction qui ne risque pas d'être déprécié par un manque de pertinence du modèle (l'estimateur pseudo direct a un biais faible, et ici une faible variance puisque n_d est grand) : c'est bien ce à quoi on aboutit puisque γ_d est grand, proche de 1.

Ajoutons qu'avec cette théorie, on est en mesure de prédire facilement chaque effet local τ_d. Après des calculs simples mais néanmoins fastidieux, on obtient :

$$\tilde{\tau}_d = \gamma_d \left(\bar{y}_d - \bar{x}_d \tilde{\beta} \right) \tag{12.14}$$

ce qui permet d'écrire l'estimateur de Henderson sous une forme plus intuitive

$$\tilde{\bar{Y}}_d^H = \bar{X}_d^T \tilde{\beta} + \tilde{\tau}_d. \tag{12.15}$$

Il reste encore une étape à franchir pour atteindre le stade opérationnel. En effet, l'estimateur BLUP $\tilde{\bar{Y}}_d^H$ a une expression complexe qui dépend à ce stade de certaines composantes du vecteur des paramètres du modèle δ introduit au 12.1.2. L'application de la stratégie BLUP a permis de produire des estimateurs $\tilde{\beta}$ de β qui ont réduit la dimension du problème : le vecteur de paramètres initiaux δ s'avère désormais limité aux composantes de variance, c'est-à-dire aux deux valeurs réelles σ_τ^2 et σ_e^2. On les résumera par le vecteur $\Sigma = (\sigma_\tau^2, \sigma_e^2)$. De fait, ce que l'on appelle - communément mais abusivement - l'estimateur $\tilde{\bar{Y}}_d^H$ n'en est pas un puisque cette expression n'est pas calculable et on devrait donc en toute rigueur le noter $\tilde{\bar{Y}}_d^H(\Sigma)$ et parler de "pseudo estimateur". Les composantes de Σ étant inconnues, il va falloir les estimer au moyen des données collectées. Une fois le paramètre Σ estimé par $\hat{\Sigma}$, on substituera $\hat{\Sigma}$ à Σ dans $\tilde{\bar{Y}}_d^H(\Sigma)$ pour aboutir à une nouvelle expression, soit $\tilde{\bar{Y}}_d^H(\hat{\Sigma})$, qui cette fois mérite bien le nom d'estimateur puisqu'elle est calculable. On donne à l'estimateur/prédicteur ainsi obtenu le nom de *Empirical Best Linear Unbiased Predictor* (EBLUP).

On estime fréquemment Σ par la méthode du maximum de vraisemblance. On dispose aussi d'une variante appelée maximum de vraisemblance restreint, qui est recommandable car elle réduit les biais des estimateurs lorsque les tailles d'échantillon sont modestes. Cette approche impose néanmoins une hypothèse supplémentaire sur la loi des variables aléatoires τ_d et e_i, que l'on considère presque systématiquement comme des variables suivant une loi de Gauss. Il n'existe pas d'expressions analytiques donnant $\hat{\sigma}_\tau^2$ et $\hat{\sigma}_e^2$, mais des algorithmes d'analyse numérique sont capables de produire des estimations conformes à la théorie. Partant de ces estimations, on obtient $\hat{\gamma}_d = \dfrac{\hat{\sigma}_\tau^2}{\hat{\sigma}_\tau^2 + \frac{\hat{\sigma}_e^2}{n_d}}$, puis :

$$\hat{\beta} = \left(\sum_{d=1}^{D} \left(\sum_{i \in s \cap d} X_i X_i^T - \hat{\gamma}_d n_d \bar{x}_d \bar{x}_d^T \right) \right)^{-1} \left(\sum_{d=1}^{D} \left(\sum_{i \in s \cap d} X_i Y_i - \hat{\gamma}_d n_d \bar{x}_d \bar{y}_d \right) \right) \tag{12.16}$$

et finalement l'estimateur EBLUP :

$$\hat{\bar{Y}}_d^H = \hat{\gamma}_d [\bar{y}_d + (\bar{X}_d - \bar{x}_d)^T \hat{\beta}] + (1 - \hat{\gamma}_d) \bar{X}_d^T \hat{\beta}. \tag{12.17}$$

Noter qu'on peut éviter toute hypothèse portant sur la loi de Y_i en utilisant une méthode de type "méthode des moments", mais en contrepartie elle s'avère théoriquement moins efficace si la distribution des variables aléatoires τ_d et e_i est effectivement gaussienne.

12.2.2 Application au modèle linéaire individuel avec corrélation spatiale

La stratégie BLUP, avec son prolongement naturel qu'est l'EBLUP, s'applique exactement de la même façon dès lors que l'on introduit des corrélations spatiales entre les effets locaux. La différence avec le modèle linéaire standard réside uniquement dans les expressions mathématiques des différents estimateurs, qui sont évidemment beaucoup plus compliquées, mais les principes ne changent pas. Détailler l'expression formelle de l'estimateur de Henderson en présence de corrélations spatiales ne peut raisonnablement se faire qu'en utilisant des notations matricielles, qui sont très lourdes et sans valeur ajoutée didactique.

L'estimateur BLUP (ou EBLUP) reste une combinaison d'un estimateur direct et d'un estimateur synthétique, avec une pondération optimale calculée en tenant compte du contexte, selon la confiance que l'on peut accorder au modèle et selon la taille de l'échantillon répondant n_d. Le coefficient σ_τ^2 introduit dans l'équation 12.3 conserve un rôle essentiel, mais les calculs doivent désormais se faire en prenant également en compte le coefficient supplémentaire ρ, lequel règle l'intensité de la corrélation spatiale. Le paramètre de modèle à estimer est donc $\Sigma = (\rho, \sigma_\tau^2, \sigma_e^2)$.

Les algorithmes de calcul du maximum de vraisemblance (restreint le cas échéant) s'accommodent de l'introduction d'un paramètre supplémentaire, et ils produisent une estimation de ρ, de σ_τ^2 et de σ_e^2. La complexité de la structure de variances-covariances ne semble pas autoriser d'autres méthodes d'estimation de Σ que celle du maximum de vraisemblance ou du maximum de vraisemblance restreint.

12.2.3 Application au modèle de Fay et Herriot

Le modèle de Fay et Herriot revêt une grande importance car il est très utilisé en pratique. Dans de nombreuses circonstances, il s'ajuste bien et produit des estimations satisfaisantes, préférables aux estimations directes. Bien que l'on se place à un degré d'agrégation plus élevé que dans les modèles précédents, la stratégie BLUP se décline aussi dans le cadre de ce modèle. Dans l'expression de l'estimateur optimum de Henderson avec le modèle standard, les σ_e^2 ont évidemment disparu mais on trouve en revanche les valeurs des vraies variances d'échantillonnage par domaine ψ_d. Il est important de noter que dans la théorie standard, les vraies variances d'échantillonnage sont supposées connues. Ce n'est évidemment pas le cas en réalité, et il faut *in fine* remplacer les expressions théoriques ψ_d par les estimateurs $\hat{\psi}_d$ que l'on obtient en appliquant les méthodes traditionnelles de calcul de variance d'échantillonnage. À ce stade, il est conseillé de terminer par un lissage des valeurs $\hat{\psi}_d$. Cette opération protège contre la prise en compte d'estimations $\hat{\psi}_d$ anormalement faibles ou anormalement fortes, évitant ainsi un impact fortement dégradant sur la qualité des estimations finales par domaine. On aboutit à :

$$\tilde{\bar{Y}}_d^H = \gamma_d \hat{\bar{Y}}_d + (1-\gamma_d)\bar{X}_d^T \tilde{\beta} \tag{12.18}$$

avec

$$\begin{aligned} \gamma_d &= \frac{\sigma_\tau^2}{\sigma_\tau^2 + \hat{\psi}_d} \\ \tilde{\beta} &= \left[\sum_{d=1}^{D} \frac{\bar{X}_d . \bar{X}_d^T}{\sigma_\tau^2 + \hat{\psi}_d}\right]^{-1} \cdot \left[\sum_{d=1}^{D} \frac{\bar{X}_d . \hat{\bar{Y}}_d}{\sigma_\tau^2 + \hat{\psi}_d}\right]. \end{aligned} \tag{12.19}$$

L'estimateur $\tilde{\bar{Y}}_d^H$ conserve une forme composite et la stratégie BLUP produit la pondération γ_d idéale, partagée entre l'estimation directe $\hat{\bar{Y}}_d$, indépendante du modèle mais instable, et l'estimation synthétique $\bar{X}_d^T \tilde{\beta}$, qui est pour sa part totalement dépendante du modèle, mais en contrepartie peu

sensible à la composition de l'échantillon répondant. Dans de rares circonstances, lorsqu'il s'agit d'estimer une proportion, il peut arriver que l'estimation $\tilde{\bar{Y}}_d^H$ sorte de l'intervalle $[0,1]$. Dans ce cas, il est nécessaire d'adapter le modèle initial.

Si on introduit des corrélations spatiales, les expressions ci-dessus évoluent en conséquence - en se compliquant considérablement - mais aucun des grands principes n'est modifié. Dans tous les cas, avec ou sans corrélations spatiales, le logiciel sait produire l'estimateur σ_τ^2 par maximum de vraisemblance (éventuellement restreint), dont on déduit immédiatement $\hat{\gamma}_d$ et $\hat{\beta}$, puis l'estimateur final EBLUP $\hat{\bar{Y}}_d^H$. Noter qu'en l'absence de corrélation spatiale, il existe d'autres méthodes d'estimation du paramètre σ_τ^2 que le maximum de vraisemblance.

12.2.4 Stratégie pour les modèles non linéaires

Le monde des modèles non linéaires est techniquement beaucoup plus compliqué que celui des modèles linéaires. En particulier, la stratégie BLUP n'est pas directement adaptée à ce contexte parce qu'elle ne trouve pas de solution mathématique satisfaisante. Elle reste néanmoins une technique de base et c'est pourquoi l'une des façons de traiter les modèles non linéaires, comme le modèle logistique ou le modèle de Poisson par exemple, consiste à les remplacer par des modèles approchés ayant une structure linéaire. Ce qu'est un modèle approché renvoie à une théorie compliquée mais néanmoins opérationnelle. C'est en partant du modèle linéaire approché que l'on applique la stratégie BLUP.

Le modèle d'origine utilise ou non des corrélations spatiales. Au modèle linéaire approché, on applique alors les développements présentés dans les parties qui précèdent.

Cela étant, l'approche la plus convaincante consiste à utiliser une stratégie mieux adaptée à ce contexte non linéaire, comme la stratégie *Empirical Bayes* qui produit des estimations optimales, ou la stratégie *Hierarchical Bayes* qui correspond à l'approche bayésienne classique.

12.3 La qualité des estimateurs

L'approche par modèle aura pour conséquence bien évidente de faire dépendre l'estimation du choix du modèle et se posera donc la question de la pertinence du modèle retenu. En effet, la simplification a un coût en termes de qualité et on peut se demander jusqu'à quel point ce modèle est correctement représentatif de la réalité.

De quoi parle-t-on ?

En matière d'appréciation de la qualité des estimations sur petits domaines, il est plus que jamais nécessaire de préciser le concept de qualité. En effet, le contexte souffre d'une complication toute particulière due à la coexistence d'aléas de natures différentes : d'une part l'aléa de sondage qui décide de la composition de l'échantillon, d'autre part l'aléa du modèle qui traite la variable d'intérêt comme une variable aléatoire. Or on peut apprécier la qualité en prenant en compte ou non l'aléa de modèle.

Sans aléa de modèle, il s'agit de l'approche classique du statisticien d'enquête placé en population finie et traitant de variables individuelles déterministes. De ce point de vue, la situation est extrêmement simple : tous les estimateurs "petits domaines" présentés jusqu'ici sont biaisés. C'est la conséquence naturelle de l'absence de prise en compte des poids de sondage (lorsque l'échantillonnage n'est pas à probabilités égales en tout cas), ou d'une prise en compte seulement partielle de ces poids. Par exemple, dans le modèle linéaire standard individuel, la pondération reflétant l'échantillonnage est systématiquement absente. Dans le modèle de Fay et Herriot, on la retrouve certes dans la composante directe $\hat{\bar{Y}}_d$ mais aucunement dans la partie synthétique $\bar{X}_d^T \hat{\beta}$. En revanche, le modèle apporte un avantage déterminant en termes de variance d'échantillonnage : en effet, les paramètres β qui sont estimés mobilisent l'intégralité de l'échantillon répondant et

c'est pourquoi ils ont une faible variance d'échantillonnage. L'effet aléatoire local estimé $\hat{\tau}_d$ est pour sa part instable, mais si le modèle est bien adapté, il sera numériquement petit et donc sa variance sera d'influence limitée. L'estimateur de Henderson devrait finalement être de variance d'échantillonnage limitée et *a priori* inférieure à celle de l'estimateur direct si le modèle a un bon pouvoir explicatif.

Lorsqu'on prend en compte l'aléa de modèle, si le modèle est linéaire, par construction l'estimateur BLUP est sans biais. Le passage à l'EBLUP n'occasionne que des biais négligeables. Si le modèle n'est pas linéaire, le contexte est beaucoup plus compliqué mais on s'attend à obtenir des biais modestes.

12.3.1 Un processus itératif

L'appréciation de la qualité peut se concevoir selon un mécanisme cyclique (figure 12.1).

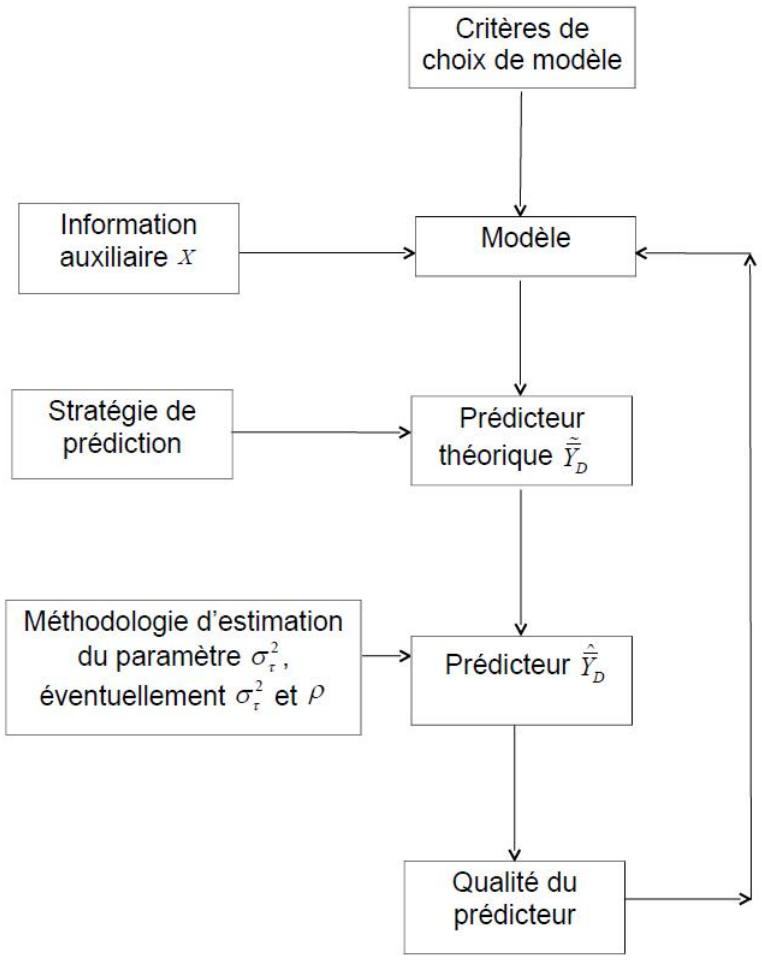


FIGURE 12.1 – Schéma du processus itératif d'appréciation de la qualité des estimateurs

Disposant d'une part de certains critères de sélection de variables explicatives, disposant d'autre part d'un ensemble de variables auxiliaires X potentiellement explicatives de Y, on ajuste un modèle. À ce stade, on dispose d'outils statistiques pour évaluer la qualité de cet ajustement. Ce modèle, associé à une stratégie de prédiction, produit un estimateur théorique $\tilde{\bar{Y}}_D$. Ce dernier est dépendant de paramètres participant à la définition du modèle (au moins le paramètre σ_τ^2, le σ_ε^2 s'il y a lieu et

le ρ s'il y a corrélation spatiale). Ces paramètres sont estimés par une méthode *ad hoc*. À la fin du cycle, on évalue la qualité du prédicteur final (biais, MSE ; voir sections 12.3.3 et 12.3.4). Si elle n'est pas acceptable, on enclenche un nouveau cycle en s'interrogeant de nouveau sur la pertinence du modèle, voire sur celle de la stratégie de prédiction ou encore sur celle de l'estimation des paramètres du modèle. L'appréciation de la qualité passe aussi par une vérification de la pertinence des hypothèses de loi du modèle, s'il y a lieu. C'est pourquoi on vérifiera le caractère gaussien des effets locaux estimés $\hat{\tau}_d$ dès lors qu'une technique de maximum de vraisemblance (restreint ou non) a été utilisée.

12.3.2 Le problème du biais

Les statisticiens d'enquête ont parfois des réticences à utiliser un estimateur dépendant d'un modèle (bien que ce soit incontournable pour traiter la non-réponse). Leur crainte essentielle est celle d'un biais substantiel si on s'en tient à l'aléa d'échantillonnage. Ce risque est inévitable puisque le modèle simplifie, et donc dénature, la réalité. L'important n'est pas d'échapper au biais mais d'obtenir un biais limité qui soit plus que compensé par le gain en termes de variance. Sauf si on travaille sur des populations artificielles, les calculs de biais dûs à l'échantillonnage ne sont pas réalisables, mais on peut utiliser deux outils simples qui permettent d'apprécier la situation, sans toutefois fournir la moindre preuve.

Le premier outil est purement graphique et consiste à construire un nuage de points où chaque point représente un des D domaines traités. Sur l'un des axes, on porte l'estimation directe (donc obtenue sans modèle), sur l'autre axe on porte l'estimation "petits domaines" (donc issue d'un modèle). Si le nuage de points ainsi formé n'est pas symétrique par rapport à la droite $y = x$ (première bissectrice), on soupçonne fortement un biais dû à l'échantillonnage. Néanmoins, il n'y a pas de fatalité (penser à la situation, évidemment idéalisée, d'un modèle traduisant une réalité où toutes les moyennes par domaine sont égales). La situation réciproque est plus convaincante : si le nuage de points est symétrique, il est probable qu'il n'y aura pas de biais significatif dû à l'échantillonnage. Le plus souvent, en pratique on observe un nuage incliné par rapport à la première bissectrice et dont la projection sur l'axe représentant l'estimation "petits domaines" est plus réduite que la projection sur l'axe représentant l'estimation directe. Ce phénomène a reçu le nom de *shrinkage*, et il est donc plutôt annonciateur d'un biais dû à l'échantillonnage. Il traduit une forme de concentration (*a priori* excessive) des estimations. Elle découle mécaniquement du modèle simplificateur, qui a un effet de normalisation et qui a donc plus ou moins tendance à uniformiser les estimations par domaine. Nous insistons sur le fait que cette approche graphique n'offre aucune preuve mais crée seulement des suspicions. En pratique, parce qu'elle ne peut pas traduire fidèlement la réalité, toute modélisation crée fatalement un biais théorique dû à l'échantillonnage et la symétrie éventuelle du nuage de points ne fait qu'indiquer le caractère probablement modeste de ce biais.

La seconde technique est encore plus simple et plus intuitive : il s'agit de sommer les estimations des totaux $\hat{T}_d^H$ obtenues sur les D petits domaines et de comparer le résultat à l'estimation directe du total $\hat{T}$ portant sur la population complète, celle qui résulte de la théorie classique des sondages. En effet, cette dernière est par construction sans biais (l'aléa est ici exclusivement l'aléa de sondage) : s'il existe un biais dû au modèle et que ce biais a un caractère quelque peu systématique, on constatera un écart entre les deux valeurs. En revanche, un biais sans composante systématique n'est pas détectable puisque des compensations peuvent se produire au moment de la sommation.

On a coutume d'exploiter, au profit de la qualité, l'écart dont il est question ci-dessus. En effet, si $\hat{T}_d^H$ est l'estimateur "petits domaines" du vrai total T_d dans le domaine d, si $\hat{T}$ est l'estimateur sans biais direct issu de l'échantillon global répondant s représentant la population complète U, on

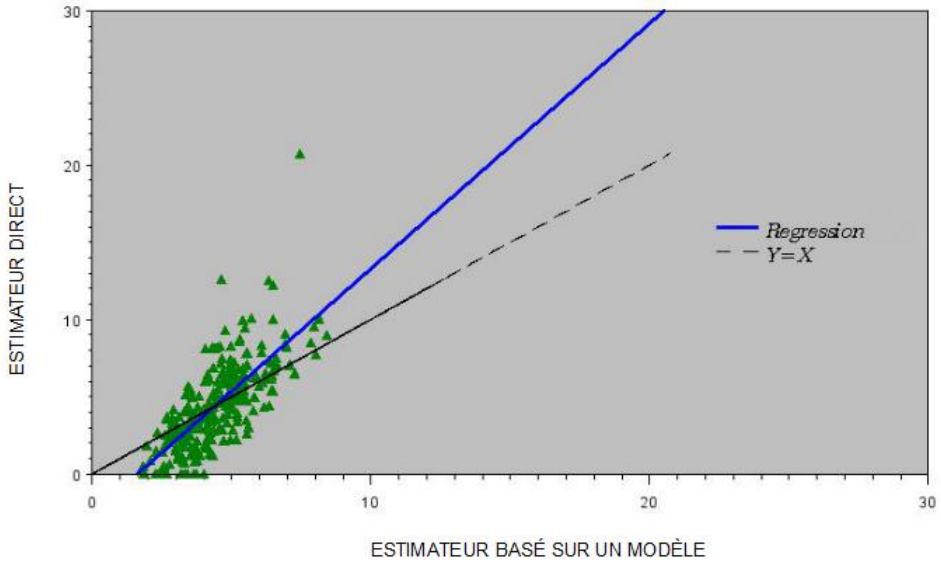


FIGURE 12.2 – Exemple de relation entre l'estimation directe et l'estimation petits domaines

adopte très souvent l'estimation finale suivante :

$$\hat{\hat{T}}_d^H = \hat{T}_d^H \frac{\hat{T}}{\sum_{d=1}^{D} \hat{T}_d^H} \tag{12.20}$$

qui permet de caler sur $\hat{T}$ l'estimation du total dans U. Cette opération reçoit le nom de *benchmarking* et contribue à limiter le biais de $\hat{\hat{T}}_d^H$ tout en assurant une diffusion cohérente.

Par ailleurs, il est toujours intéressant de procéder à une cartographie des estimations de moyenne par domaine $\hat{\hat{Y}}_d^H$, laquelle permet de vérifier visuellement la cohérence du système d'estimation dans son ensemble : normalement, à deux domaines ayant des caractéristiques proches et voisins sur une carte, devraient correspondre deux moyennes estimées $\hat{\hat{Y}}_d^H$ semblables (concrètement, les couleurs représentatives de leurs valeurs respectives devraient se situer dans la même gamme).

12.3.3 L'erreur quadratique moyenne

Dans un environnement où les biais sont possibles, ou probables, ou encore inévitables, le bon concept d'erreur est celui d'erreur quadratique moyenne (ou MSE). Cet indicateur désigne l'espérance du carré de la différence entre l'estimateur et le paramètre. En prenant en compte à la fois l'aléa de sondage et l'aléa de modèle, le cadre théorique offert par le modèle permet d'obtenir l'expression de la MSE et ensuite de l'estimer sans biais ou presque. L'expression de la MSE et son estimation sont très compliquées, même avec le modèle linéaire, et le calcul est donc confié à un logiciel. Néanmoins, en l'absence de corrélation spatiale, on peut vérifier, si le nombre de domaines D est grand, que le terme numériquement le plus important dans l'estimation de la MSE de $\hat{\hat{Y}}_d^H$ est $\hat{\gamma}_d \hat{\psi}_d$ pour le modèle de Fay et Herriot, et $\hat{\gamma}_d \frac{\hat{\sigma}_e^2}{n_d}$ pour le modèle linéaire individuel standard. Concernant tous ces calculs d'erreur, les résultats obtenus supposent fondamentalement que le modèle est spécifié de manière parfaitement conforme à la réalité (le modèle peut être qualifié d'"exact"). Cela n'est certainement pas vrai en toute rigueur ! L'introduction d'une corrélation spatiale crée évidemment une difficulté technique supplémentaire, mais la théorie générale permet

d'aboutir, ce qui ne signifie pas que les outils informatiques actuellement accessibles soient en capacité de la mettre en œuvre. Notez que dans certaines circonstances particulièrement favorables, on peut disposer d'une source externe qui fournit la vraie valeur du paramètre (par exemple à l'issue d'un recensement). Cela permet d'apprécier directement l'erreur d'estimation commise.

12.4 Mise en œuvre avec R

■ **Exemple 12.1 — Diffusion du recensement sur des carreaux.** L'Office statistique de l'Union européenne EUROSTAT souhaite produire en 2021 des statistiques (sexe, tranche d'âge, activité, etc.) portant sur la population complète de chaque pays membre au niveau de carreaux d'un kilomètre de côté. De surcroît, en France, l'Insee a pour ambition de diffuser les données du Recensement de la Population (RP) sur des carreaux de quelques centaines de mètres de côté. Depuis 2004, le recensement de la France est effectué par sondage dans les communes de plus de 10 000 habitants [2]. Dès lors, la superficie ciblée contient trop peu d'observations pour que l'on puisse obtenir de bons estimateurs directs des paramètres d'intérêt. C'est la raison pour laquelle l'estimation *petits domaines* pourrait être une technique statistique appropriée pour l'exploitation de ce type de données.

L'introduction d'une corrélation spatiale dans ce contexte permet de traduire le phénomène de continuité naturelle des caractéristiques sociodémographiques d'individus peuplant des zones géographiquement contiguës. En effet, passant d'un carreau quelconque aux carreaux voisins, on ne peut pas raisonnablement prétendre *a priori* qu'il y ait indépendance entre les comportements des unités statistiques – logements comme individus – qui les composent. ■

Le package *sae* de R permet de calculer les estimations petits domaines aux niveaux "domaine" et "individu", dans le cas de modèles respectivement sans et avec prise en compte de l'autocorrélation spatiale. Implémenté par Molina et Marhuenda, ce package a fait l'objet d'une présentation dans *The R Journal* (MOLINA et al. 2015).

Les principales fonctions qui ont été utilisées pour traiter les données du RP à partir d'un modèle formulé au niveau domaine sont issues du package *sae* : il s'agit de `eblupFH()`, `eblupSFH()`, `mseFH()` et `mseSFH()`.

Pour produire des estimations à partir d'un modèle au niveau individu, on a utilisé les fonctions `eblupBHF()` et `pbmseBHF()` du package *sae*, ainsi que la fonction `corrHLfit()` du package *spaMM*.

Modélisation au niveau domaine : les fonctions de base eblupFH() et mseFH()

Les premières estimations s'appuient sur le modèle de Fay et Herriot sans autocorrélation spatiale. La fonction `eblupFH()` permet d'obtenir en sortie :

i) les estimations de Fay et Herriot pour chaque domaine ;
ii) une estimation de la variance σ_τ^2 de l'effet aléatoire propre aux domaines.

En supplément, la fonction `mseFH()` produit le calcul des erreurs quadratiques moyennes associées à chaque estimation (voir partie 12.3.3).

Les arguments de ces fonctions sont les mêmes. La syntaxe standard est la suivante :

```
mod_FH <- eblupFH(formula = Y ~ X1+...+Xp, vardir = varech, method =  ,
   maxiter = m, precision = e, B = 0, data = )
```

En premier lieu, le paramètre `formula` précise la variable d'intérêt Y ainsi que les variables explicatives retenues $X_1, ..., X_p$. Les valeurs numériques de toutes ces variables doivent être contenues dans un tableau qui associe une ligne à chaque domaine, précisé dans l'argument `data`. Les

2. Il est exhaustif dans les communes de moins de 10 000 habitants

paramètres $\hat{\gamma}_d$ impliqués dans les estimateurs de Henderson sont calculés à l'aide des variances d'échantillonnage (estimées) par domaine $\hat{\psi}_d$, disponibles dans une variable que l'on précise dans l'argument `vardir`.

L'estimation de l'unique paramètre de variance du modèle est obtenue par une technique *ad hoc*. Concrètement, on utilise une méthode itérative qui devrait converger vers la valeur σ_τ^2. Les paramètres `maxiter` et `precision` sont des paramètres techniques (définis soit par l'utilisateur soit par défaut) qui régulent ce processus itératif. À chaque itération, l'algorithme calcule une estimation de σ_τ^2. Le rôle du paramètre `precision` est le suivant : dès que la différence entre deux valeurs obtenues consécutivement est inférieure à celui-ci (e dans notre exemple), l'algorithme s'arrête. Sinon, tant que le nombre maximal d'itérations `maxiter` n'est pas atteint, les itérations se poursuivent. La sortie indique si l'algorithme a convergé ou non. La méthode est également à préciser. On peut choisir parmi trois méthodes, dont la méthode du maximum de vraisemblance et celle du maximum de vraisemblance restreint (respectivement `method = "ML"` et `method = "REML"`). La troisième méthode (`method = "FH"`) est une méthode de type méthode des moments.

On attire l'attention du lecteur sur la nécessité de ne pas rajouter dans les régresseurs de variables indicatrices repérant les domaines. En effet, la constante faisant déjà partie des régresseurs standards, cette pratique conduirait à former une matrice non inversible. La commande suivante conduit donc à un échec :

```
mod_FH <- eblupFH(formula = Y ~ X1+...+Xp+as.factor(Carreau) , vardir =
    varech, method =  , maxiter = m, precision = e, B = 0, data = )
```

Cas d'une corrélation spatiale au niveau domaine : les fonctions eblupSFH() et mseSFH()

Le logiciel estime un modèle SAR (voir partie 12.1.3) du type :

$$\tau = \rho.A.\tau + u. \tag{12.21}$$

Les paramètres de ces fonctions sont les mêmes que ceux des fonctions précédentes, à cela près que l'on doit en plus préciser la matrice de proximité $\mathbf{A}$ (la matrice des coefficients $\alpha_{i,j}$; voir partie 12.1.3). La commande R est de la forme :

```
mod_SFH <- eblupSFH(formula = Y ~ X1+...+Xp, vardir = varech, method =  ,
    maxiter = m, precision = e, proxmat = A, B = 0, data = )
```

La matrice de proximité est décrite par le paramètre `proxmat`. Elle a des lignes standardisées en ce sens où la somme des éléments de chaque ligne vaut toujours 1.

Un processus itératif analogue à celui du cas sans corrélation spatiale permet de calculer :

i) les estimations de Fay et Herriot pour chaque domaine ;
ii) une estimation de la variance de l'effet aléatoire propre aux domaines ;
iii) une estimation du paramètre d'autocorrélation spatiale ρ.

Modélisation au niveau individu, sans corrélation spatiale : les fonctions eblupBHF() et pbmseBHF()

En utilisant une modélisation au niveau individu, la fonction `eblupBHF()` du package *sae* permet de calculer les estimations directes et les estimations "petits domaines" sans corrélation spatiale. La syntaxe est la suivante :

```
mod_BHF <- eblupBHF(formula = Y ~ X1+...+Xp, dom = ,
meanxpop = , popnsize = Popn, data = adr_est)
```

Elle utilise le paramètrage suivant : `formula` pour l'expression formelle du modèle, `dom` pour désigner la variable identifiant les domaines, `popnsize` pour la taille de la population N_d dans chaque domaine, et `meanxpop` pour les moyennes des variables explicatives $\bar{X}_d$ calculées dans la population complète du domaine. Le paramètre `data` désigne la table des données.

La fonction `pbmseBHF()` estime les erreurs (MSE) des estimateurs "petits domaines" par une technique de *bootstrap*. Les paramètres de cette fonction sont les mêmes que pour la fonction précédente, avec en supplément le nombre de ré-échantillonnages du *bootstrap* défini par le paramètre `B` (B=1000 par exemple).

```
mse_BHF <- pbmseBHF(formula = Y ~ X1+...+Xp, dom = ,
meanxpop = , popnsize = , B = 1000, data = )
```

Prise en compte de la corrélation spatiale dans le modèle individuel

Le package *spaMM* peut être utilisé pour prendre en compte la corrélation spatiale. Il peut gérer plusieurs types de modèles, et en particulier le modèle de Poisson (voir partie 12.1.4). La fonction `corrHLfit()` traite le modèle de Poisson au niveau individu avec corrélation spatiale.

```
library(spaMM)
mod_spa <- corrHLfit(formula = Y ~ X1+...+Xp+Matern(1|x+y),
HLmethod = "REML", family = "poisson", ranFix = list(nu=0.5), data = )
```

Concernant le paramètrage de cette fonction, `formula` désigne l'expression formelle du modèle. La composante *Matern(1|x+y)*, propre à la fonction utilisée, permet de prendre en compte les coordonnées x et y des domaines (ici les centres des carreaux), qui doivent donc être présentes dans la table des données, afin de calculer les distances qui interviennent dans la fonction de corrélation spatiale. Par ailleurs, `HLmethod` précise la méthode d'estimation des paramètres de variance et de corrélation spatiale (ici le maximum de vraisemblance restreint), et `family` choisit la distribution de la variable d'intérêt (ici une loi de Poisson). La forme fonctionnelle de la corrélation spatiale peut être choisie parmi une famille paramétrée de fonctions compliquées appelées fonctions de Matérn. Le paramètre `ranFix` précise le paramétrage de cette famille de fonctions. Si on indique `list(nu=0.5)`, on obtient la forme exponentielle de l'équation 12.3, qui est l'expression traditionnellement utilisée - à cela près que le paramètre estimé est $\frac{1}{\rho}$ et non directement ρ. Le paramètre `data` désigne la table des données.

On obtient en sortie, entre autres, les coefficients estimés du modèle, dont le coefficient ρ intervenant dans la corrélation spatiale exponentielle (en fait son inverse si on se réfère à la formulation 12.3), les prédictions optimales $\hat{\tau}_d$ des effets locaux aléatoires, et la variance estimée de l'effet aléatoire $\hat{\sigma}^2_\tau$.

Conclusion

L'estimation sur petits domaines est fondée sur l'utilisation de modèles stochastiques. C'est la contrepartie d'une certaine pauvreté de l'information collectée *via* l'échantillon recoupant le domaine lorsque celui-ci est de petite taille. En effet, pour limiter l'imprécision, il n'y a pas de miracle, il faut bien formuler des hypothèses portant sur l'intégralité de la population et qui compensent le manque d'information obtenue au niveau local. Les modèles font intervenir explicitement des effets géographiques locaux dont l'interprétation est délicate en ce sens qu'on peut toujours considérer qu'il s'agit d'un pis-aller pour camoufler une prise en compte insuffisante d'effets fixes explicatifs du phénomène étudié. Au fond, la première question est bien de savoir jusqu'à quel point l'effet purement géographique existe. Par ailleurs, ces modèles, quels qu'ils soient, créent toujours du biais par rapport à l'aléa de sondage. L'objectif essentiel est d'en limiter l'ampleur, plus que de mesurer la variance d'échantillonnage, qui devient un objectif secondaire pour le statisticien d'enquête. On

dispose certes d'outils statistiques pour apprécier la qualité de l'ajustement d'un modèle, mais cela ne garantit pas l'adéquation du modèle retenu à la situation particulière d'un domaine donné, qui peut être très spécifique sans que le statisticien ne s'en aperçoive. Il n'existe pas d'estimation fiable du biais d'échantillonnage, et on ne dispose actuellement que de quelques outils qualitatifs, plus ou moins convaincants et qui conduisent seulement à l'appréciation d'une situation globale. De façon générale, la théorie des modèles linéaires (*Linear Mixed Models* ou LMM) est beaucoup plus simple que celle des modèles non linéaires traditionnellement mis en œuvre (*Generalized Linear Mixed Models* ou GLMM), qui restent vraiment difficiles d'accès. La présence de corrélation spatiale complique toujours le contexte et se pose alors la question de la disponibilité du code informatique pour procéder aux estimations. Le développement de R est prometteur et on devrait aller à l'avenir vers un élargissement de la gamme des modèles acceptant de la corrélation spatiale.

Références - Chapitre 12

BATTESE, George E, Rachel M HARTER et Wayne A FULLER (1988). « An error-components model for prediction of county crop areas using survey and satellite data ». *Journal of the American Statistical Association* 83.401, p. 28–36.

CHANDRA, Hukum, Ray CHAMBERS et Nicola SALVATI (2012). « Small area estimation of proportions in business surveys ». *Journal of Statistical Computation and Simulation* 82.6, p. 783–795.

COELHO, Pedro S et Luıs N PEREIRA (2011). « A spatial unit level model for small area estimation ». *REVSTAT–Statistical Journal* 9.2, p. 155–180.

FAY III, Robert E et Roger A HERRIOT (1979). « Estimates of income for small places : an application of James-Stein procedures to census data ». *Journal of the American Statistical Association* 74.366a, p. 269–277.

MOLINA, Isabel et Yolanda MARHUENDA (2015). « sae : An R package for small area estimation ». *R Journal, in print*.

PRATESI, Monica et Nicola SALVATI (2008). « Small area estimation : the EBLUP estimator based on spatially correlated random area effects ». *Statistical methods and applications* 17.1, p. 113–141.

RAO, John NK (2015). *Small-Area Estimation*. Wiley Online Library.

IV Partie 4 : Prolongements

13. Partitionnement et analyse de graphes

PASCAL EUSEBIO, JEAN MICHEL FLOCH, DAVID LEVY
Insee

Résumé

Analyser le réseau des villes a nécessité de s'éloigner des méthodes habituellement utilisées à l'Insee et de recourir à des représentations sous forme de graphes. Si ces techniques sont encore peu répandues dans la statistique publique, le problème posé est assez classique : réaliser la partition d'une population en sous-populations. On cherche à repérer des sous-populations homogènes (faible hétérogénéité intra-classe) et assez différenciées (forte hétérogénéité inter-classe). En utilisant les graphes, nous verrons que nous recherchons souvent des partitions qui conservent beaucoup de flux intra-zones et peu de flux entre elles. Les solutions algorithmiques reposent sur des méthodes "agglomératives" ou "divisives" selon les cas, que nous pouvons rapprocher des méthodes ascendantes ou descendantes que nous connaissons en analyse des données. Elles utilisent la notion de modularité, fondée sur le comparaison du graphe étudié à un graphe aléatoire.

Ce chapitre n'a pas vocation à balayer l'ensemble des méthodes de la théorie des graphes qui ont connu de fortes évolutions depuis leur apparition dans les années 1930. Ces méthodes ont été élaborées dans des domaines très divers (géographie, analyse des réseaux sociaux, biologie, informatique). Les méthodes présentées sont issues essentiellement du monde de la physique (autour du concept clé de modularité) mais un encadré fournit quelques compléments sur les méthodes de *blockmodelling* et sur la prise en compte de l'espace dans les réseaux.

13.1 Les graphes et l'analyse géographique des réseaux de villes

Les géographes se sont intéressés depuis longtemps à l'analyse des relations entre territoires. De nombreux travaux ont porté sur les hiérarchies urbaines. On peut citer parmi les exemples anciens la théorie des lieux centraux de CHRISTALLER 2005. Les données disponibles et les outils de traitement ont longtemps limité l'analyse des flux. Les modèles gravitaires issus des travaux de Wilson ont constitué une façon simple de modéliser les interactions (WILSON 1974). C'est avec des développements spécifiques de la théorie des graphes, issus d'autres domaines que celui de la géographie, que la situation a été considérablement modifiée (sociologie pour quelques intuitions, physique, informatique). Deux modèles de graphes ont eu une importance particulière : les graphes petit-monde et les graphes invariants d'échelle.

Définition 13.1.1 — Graphe. Un **graphe** est une représentation graphique d'un ensemble de sommets reliés par des arêtes.
Une **arête** est un lien entre deux éléments distincts.
Un **sommet** ou **nœud** est un élément relié par des arêtes. Le **degré** d'un sommet est le nombre de sommets auquel il est relié.

■ **Exemple 13.1** Un graphe de villes représente des villes (les sommets) qui échangent des populations : les navettes domicile-travail (les arêtes, aussi appelées liens dans la suite). ■

13.1.1 Petit-monde

Pendant longtemps, les spécialistes des graphes ne s'intéressaient qu'aux graphes aléatoires encore aujourd'hui très utilisés. Dans les années 1990, divers théoriciens des graphes ont proposé des modèles comme le petit-monde et l'invariance d'échelle. Ces modèles n'ont pas été sans influence dans l'analyse géographique. Les graphes de type petit-monde ont été proposés par Watts et Strogatz dans un article de la revue *Nature* (WATTS et al. 1998). On trouve, dans la figure 13.1, la reproduction du schéma proposé par les deux auteurs pour illustrer la construction du graphe petit-monde.

Définition 13.1.2 — Graphe aléatoire. Graphe dont la distribution des arêtes est aléatoire.

L'idée du petit-monde trouve son origine (lointaine) dans les travaux de Stanley Milgram. L'expérience de Milgram consistait à demander à des habitants du Middle West de faire parvenir une lettre à un destinataire de la côte ouest, qu'ils ne connaissaient pas, en utilisant comme intermédiaires des personnes de leur entourage. Milgram eut la surprise de constater qu'en moyenne les chaînes parvenues au destinataire n'étaient composées que de 5,6 individus. Cette expérience a permis de confirmer la thèse de KARINTHY 1929 selon laquelle toutes les personnes du globe sont reliées par une chaîne d'au plus 5 maillons, devenue dans sa version populaire les six degrés de séparation : en clair, seules cinq personnes nous séparent de n'importe quelle autre personne dans le monde.

Le graphe de départ est un graphe qualifié de k-régulier.

Définition 13.1.3 — Graphe k-régulier. Graphe dans lequel chaque sommet est lié au même nombre k de sommets (BATTISTON et al. 2014). En d'autres termes, tous les sommets ont le même degré k.

L'idée des auteurs est de présenter une façon simple de transformer ce graphe régulier en graphe aléatoire. À chaque étape, un lien est supprimé de façon aléatoire avec une probabilité p, et on ajoute de la même façon un lien. Le processus est décrit de façon détaillée dans l'article fondateur. Watts et Strogatz ont combiné deux mesures : $L(p)$ et $C(p)$ pour caractériser un type de réseau.

$L(p)$ désigne la longueur moyenne du plus court chemin entre les paires de sommets lorsque

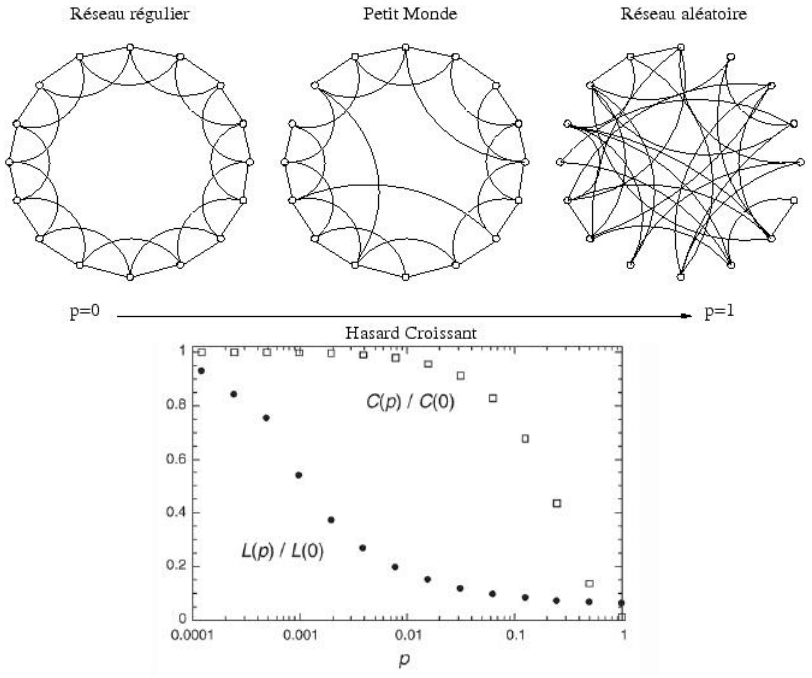


FIGURE 13.1 – Réseaux petit-monde
Source : WATTS *et al. 1998*

p varie. $C(p)$ désigne le coefficient de *clustering*, dont on trouvera une illustration dans la figure 13.2. Ce coefficient est en rapport avec la notion de transitivité dans le graphe, notion connue des sociologues depuis les années 1970. L'idée de transitivité peut être traduite de façon simple par le fait que les amis de nos amis sont souvent nos amis. Une forte transitivité dans le graphe se traduit par le fait que, du point de vue topologique, on trouve beaucoup de triangles. Strogatz et Watts ont proposé des coefficients locaux (associés) à chaque nœud du graphe, et un coefficient global, qui est la moyenne arithmétique des coefficients locaux.

Définition 13.1.4 — Le coefficient de *clustering*.

$$C(p) = \sum_{i=1}^{n} \frac{C_i}{n} \tag{13.1}$$

avec

$$C_i = \frac{\text{nombre de triangles dont un des trois sommets est le nœud i}}{\binom{k}{2}} \tag{13.2}$$

où k est le coefficient local ou le degré du nœud et n le nombre de nœuds du graphe.

■ **Exemple 13.2** Avec le graphe présenté en figure 13.2,

$$C_4 = \frac{5}{\binom{6}{2}} = 1/3$$

et le coefficient de *clustering* du réseau est :

$$C(p) = \sum_{i=1}^{n} \frac{C_i}{n} = 0,5208.$$

■

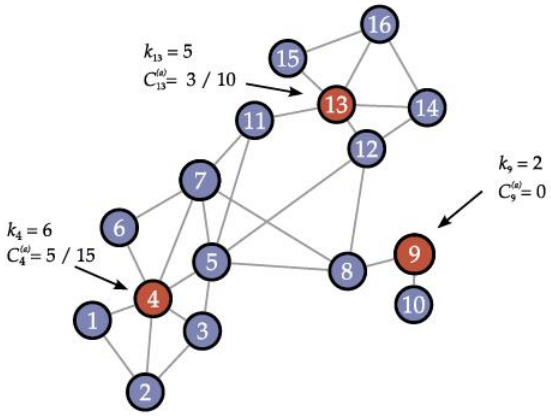


FIGURE 13.2 – Le coefficient de *clustering*

Les valeurs de $C(p)$ et $L(p)$ sont normées par les valeurs $C(0)$ et $L(0)$ correspondant à un graphe régulier. Les deux indicateurs évoluent de façon très différente. La distance moyenne entre les nœuds diminue rapidement tandis que le coefficient de *clustering* (rapport du nombre de triangle sur le nombre de triplets possibles) reste stable un moment et décroît plus rapidement. Watts et Strogatz estimaient que pour des valeurs intermédiaires de p, les réseaux restaient assez hautement structurés, à l'instar des graphes réguliers, mais avec une faible longueur moyenne des chemins, comme dans les graphes aléatoires. C'est ce qu'ils ont qualifiés de graphes petit-monde, dans une définition qui reste assez largement qualitative (grand nombre de sommets, nombre de liens existants loin de la saturation, degré important de *clustering*, faible distance moyenne). Des définitions mathématiques plus précises ont été proposées mais elles sont très techniques et dépassent notre propos.

Des réseaux de type petit-monde peuvent être générés à l'aide de la fonction `sample_smallworld` du package *igraph* de R. Dans un tel réseau, on fait l'hypothèse que chaque sommet puisse être relié à n'importe quel autre.

Application avec R

```
# Package necessaire
library(igraph)

# Generation du graphe avec 100 noeuds
g <- sample_smallworld(dim = 1, size = 100, nei = 5, p = 0.05)


# Representation du graphe
plot(g, vertex.size=4,vertex.label.dist=0.5,
                vertex.color="green",
        edge.arrow.size=0.5)
```

```
# calcul des coefficients du graphe
## le coefficient  local
q=transitivity(g,type = "local")

## le coefficient global
transitivity(g,type = "average")
```

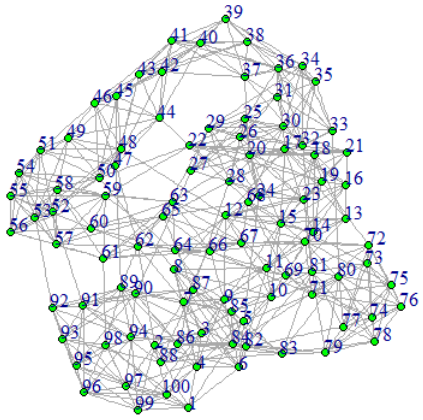


FIGURE 13.3 – Graphe petit-monde
Source : *Simulation à partir du package igraph*

```
# qui est bien egal a la moyenne des coefficients locaux
mean(q)
```

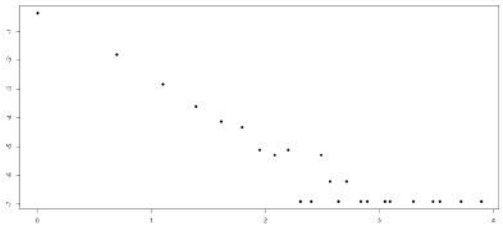
13.1.2 Réseaux invariants d'échelle

Un autre ensemble de graphes complexes est celui des graphes invariants d'échelle. Cette modélisation a été proposée initialement par BARABÁSI et al. 1999. On peut générer ce type de graphe sous R avec la fonction `barabasi.game` du package *igraph* et on en trouvera une illustration dans la figure 13.4.

La logique de constitution de ce type de graphe est notablement différente de celle des petits-mondes. Ces graphes font apparaître une distribution particulière des degrés qui est du type loi de puissance (BARABÁSI et al. 1999). Chaque nouveau nœud aura une probabilité de se lier à un nœud d'autant plus forte que le degré de ce nœud est élevé. Ils sont appelés invariants, car un zoom sur n'importe quelle partie du graphe ne change pas sa forme. À chaque niveau de grossissement, le réseau contiendra quelques nœuds avec beaucoup de connexions et un grand nombre de nœuds avec très peu de connexions. Ainsi, le réseau est dit **invariant d'échelle** si, lorsque k désigne le degré, et $P(k)$ la fréquence des sommets de degré k, l'estimation de la fonction $P(k) = k - \gamma$ fait apparaître une valeur de γ supérieure à 2. Dans l'exemple que l'on a présenté en figure 13.4, la valeur du coefficient γ est de 2,6.

Ces deux modèles, décrits ici sommairement, n'épuisent pas la description des réseaux complexes. Dans un ouvrage, Newman (auteur de plusieurs algorithmes de partitionnement de graphes), Barabasi (introducteur des graphes invariants d'échelle) et Watts (introducteur des graphes petit-monde) montrent que les graphes complexes combinent souvent des caractéristiques des deux types (NEWMAN et al. 2011). Cela est très net dans les réseaux urbains que nous allons aborder : on rencontre souvent des communautés de villes présentant de fortes interactions (caractéristiques petit-monde) tandis qu'au niveau supérieur, les liens entre communautés relèvent plutôt d'une logique invariance d'échelle. De nombreux travaux ont été menés sur les réseaux de villes. On

(a) Exemple de graphe de Barabasi



(b) Évolution de la fréquence du nombre de voisins

FIGURE 13.4 – Exemple de réseaux invariants d'échelle
Source : *Graphes simulés par la fonction barabasi.game du package igraph*

peut citer divers travaux de ROZENBLAT et al. 2013 sur les réseaux de transport aérien, sur la combinaison des transports aériens et maritimes, ou sur les liens géographiques entre les firmes multinationales. On trouvera en figure 13.5 un schéma illustrant les emboîtements entre logique petit-monde et logique invariance d'échelle.

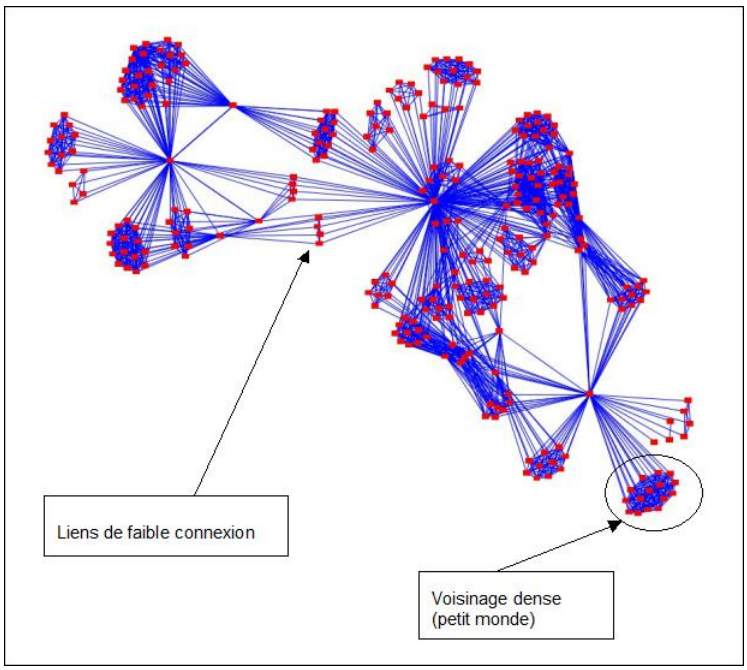


FIGURE 13.5 – Réseau formé de réseaux petit-monde et invariance d'échelle
Source : ROZENBLAT *et al. 2013*

Certains auteurs (BEAUGUITTE et al. 2011) relativisent cependant l'apport des deux concepts à la géographie, estimant que celui de petit-monde est généralement trivial tandis que celui d'inva-

riance d'échelle est connu depuis longtemps. En revanche, l'utilisation des méthodes de partitionnement, issues de travaux de physiciens, ont considérablement enrichi les possibilités d'analyse des réseaux complexes.

13.2 Les méthodes de partitionnement de graphes

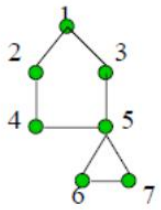
Si le graphe permet de représenter les échanges entre les sommets, le partitionnement de graphe met en évidence des groupes de sommets reliés préférentiellement. Ainsi, le partitionnement du graphe représentant les échanges commerciaux permet, par exemple, d'indiquer où implanter les plateformes de transport pour desservir au mieux le territoire : c'est au sein de chaque groupe obtenu par partitionnement. Ces méthodes sont depuis les années 2000 en plein développement, et il ne peut être question ici que d'en donner une vision introductive, en essayant de l'appuyer sur des intuitions. Elles forment une branche de la théorie des graphes, méthode assez ancienne d'analyse (problème d'Euler sur les ponts de Koenigsberg, problème de la coloration d'une carte, etc.). Les notions de la théorie des graphes classique ne seront mobilisées que lorsqu'elles seront indispensables et on se centrera sur les concepts spécifiques aux grands graphes et à leur partitionnement.

13.2.1 Notions de théorie des graphes

Définition 13.2.1 — Caractériser un graphe. Le **graphe** est un ensemble $G = \{V, E\}$ (figure 13.6) où V (de l'anglais *vertex*) désigne les sommets et E (de l'anglais *edge*) les arêtes.
La **taille** du graphe est le nombre de liens.
L'**ordre** du graphe est le nombre de sommets.
Un graphe est dit **vide** lorsqu'il ne contient aucun lien.
Un graphe est dit **complet** lorsque tous les sommets sont connectés à tous les autres. Il y a alors $\frac{n(n-1)}{2}$ liens dans un graphe complet d'ordre n.
Un graphe **orienté** est un ensemble de sommets et d'arêtes, chaque arête étant un couple de sommets ordonnés. Ainsi, la relation entre les sommets x et y est différente de celle entre y et x.
Un graphe **valué**, par opposition à un graphe non valué, comporte des liens multiples (deux sommets sont liés plusieurs fois).
Dans cette communication, on se limitera à des graphes non orientés, dans lesquelles les relations entre sommets sont de fait symétriques.
Un graphe **simple** est un graphe non valué et sans boucle (sans arête d'un sommet vers lui-même).
Le **degré** d'un sommet est le nombre de sommets auquel il est relié. Dans un graphe simple d'ordre n, le degré d'un sommet est compris ente 0 et $n-1$. La séquence des degrés est la suite $d_1, ..., d_n$.
La **densité** d'un graphe est le rapport entre le nombre de liens observés et le nombre de liens d'un graphe complet. Ainsi, elle varie entre 0 pour un graphe vide et 1 pour un graphe complet.
Si chaque point d'un graphe est atteignable depuis n'importe quel point alors le graphe est **connecté** ou **connexe**.

■ **Exemple 13.3** Le graphe présenté en figure 13.6 est un graphe simple de taille 8 et d'ordre 7. ■

Si la formalisation de la théorie devient rapidement complexe, certains concepts sont assez faciles à appréhender. Comme dans les méthodes de statistique spatiale, on peut associer au graphe une matrice d'adjacence (figure 13.7). Une valeur supérieure à 0 indique qu'il existe un lien entre deux points. Si la matrice d'adjacence est symétrique alors elle est issue d'un graphe non orienté. Si sa diagonale vaut 0 alors le graphe associé est simple (sans boucle).



(a) Un graphe à 5 nœuds et 8 arêtes

V={1,2,3,4,5,6,7}
E={(1,2),(1,3),(2,4).
(4,5),(3,5),(4,5),
(5,6),(6,7)}

(b) Son écriture mathématique

FIGURE 13.6 – Représentations géométrique et mathématique d'un graphe

0	1	1	0	0	0	0
1	0	0	1	0	0	0
1	0	0	0	1	0	0
0	1	0	0	1	0	0
0	0	1	1	0	1	1
0	0	0	0	1	0	1
0	0	0	0	1	1	0

(a) Matrice d'adjacence

2	0	0	0	0	0	0
0	2	0	0	0	0	0
0	0	2	0	0	0	0
0	0	0	2	0	0	0
0	0	0	0	4	0	0
0	0	0	0	0	2	0
0	0	0	0	0	0	2

(b) Matrice des degrés

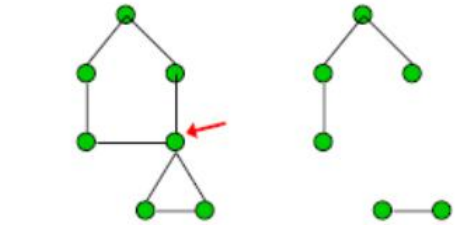
2	−1	−1	0	0	0	0
−1	2	0	−1	0	0	0
−1	0	2	0	−1	0	0
0	−1	0	2	−1	0	0
0	0	−1	−1	4	−1	−1
0	0	0	0	−1	2	−1
0	0	0	0	−1	−1	2

(c) Matrice laplacienne

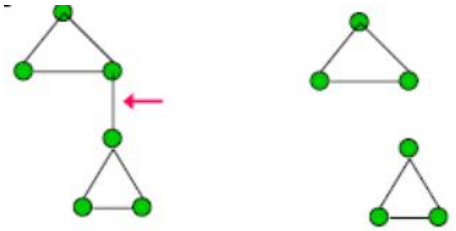
FIGURE 13.7 – Matrices d'adjacence, des degrés et laplacienne associées au graphe de la figure 13.6

Un chemin du sommet a vers le sommet b est une suite ordonnée de sommets dans laquelle chaque paire adjacente est reliée par une arête. Une **géodésique** entre deux points est le chemin de longueur minimale entre ces deux points. Dans l'exemple de la figure 13.6, la suite de sommets $(1,3,5,7)$ est la géodésique entre les points 1 et 7, et la suite de sommets $(1,2,4,5,7)$ est un chemin et non une géodésique. Un point **a** est atteignable depuis un point **b** lorsqu'il existe un chemin entre les deux points. Si on soustrait cette matrice d'adjacence à la matrice des degrés (matrice dont la diagonale est constituée des degrés de chaque sommet), on obtient la matrice **laplacienne** (figure 13.7) qui joue un rôle fondamental dans l'approche qualifiée de spectrale des graphes (méthodes de *clustering*).

Toutes les questions que l'on examinera désormais tournent autour de la possibilité de déterminer au sein de notre graphe des sous-graphes appelés **communautés** ou **cliques**. Cela conduit à s'intéresser aux sommets et aux liens qui jouent un rôle particulier, ainsi qu'aux indicateurs qui permettent de mesurer cela. Les points de coupure et les ponts renvoient respectivement aux nœuds et aux liens dont la suppression diminue la connectivité globale du graphe (figure 13.8).



(a) Point de coupure (nœud)



(b) Pont (lien)

FIGURE 13.8 – Suppression de nœuds ou de liens

Définition 13.2.2 — Quelques indicateurs de centralité. Mesures qui permettent d'appréhender les sommets (et les liens) les plus importants.
La **connectivité** d'un graphe est le nombre de sommets qu'il faut enlever pour supprimer la propriété connexe du graphe. On définit de façon duale une connectivité de liens qui correspond au nombre de liens à supprimer pour que la connectivité disparaisse.
Les indicateurs de centralité jouent un rôle très important dans l'analyse et le partitionnement d'un graphe. Plusieurs ont été définis :
La **centralité de degré** (*degree centrality*) est tout simplement le degré, c'est-à-dire le nombre de liens depuis un sommet. Dans notre exemple en figure 13.6, c'est le sommet 5 qui a la plus forte centralité de degré. Cette centralité peut être normée en la rapportant au nombre de sommets moins un. C'est la notion la plus simple. Elle est utilisée fréquemment en sociologie, mais elle ne prend pas en compte la structure du graphe.
La **centralité de proximité** (*closeness centrality*) indique si le sommet est situé à proximité de l'ensemble des sommets du graphe et s'il peut rapidement interagir avec ces sommets. Il s'écrit

formellement :

$$C_c(v) = \frac{1}{\sum_{u \in V \setminus \{v\}} d_G(u,v)} \tag{13.3}$$

avec $d_G(u,v)$ la distance entre les sommets u et v.

La **centralité d'intermédiarité** (*betweeness centrality*) est un des concepts les plus importants. Il mesure l'utilité du sommet dans la transmission de l'information au sein du réseau. Le sommet joue un rôle central si beaucoup de plus courts chemins entre deux sommets doivent emprunter ce sommet. Elle s'écrit :

$$C_B(v) = \sum_{\substack{i,j \\ i \neq j \neq v}} \frac{\sigma_{ij}(v)}{\sigma_{ij}} \tag{13.4}$$

avec $\sigma_{ij}(v)$ le nombre de chemins entre i et j qui passent par v.

Il existe aussi la centralité d'intermédiarité de lien, qui rend compte du nombre de géodésiques (plus courts chemins) qui empruntent ce lien. La figure 13.9 montre un lien (trait foncé) ayant une forte centralité d'intermédiarité. Ainsi la suppression de ce lien conduit à la formation de deux sous-graphes. Cette propriété est utilisée dans le partitionnement de graphes.

La **centralité de vecteur propre** ou **centralité spectrale** est définie par Bonacich à partir de la matrice d'adjacence. La centralité spectrale est une mesure de l'influence d'un nœud au sein d'un réseau. Elle correspond pour un sommet à la somme de ses connexions avec les autres sommets, pondérée par la centralité de degré de ces sommets. On peut l'écrire sous la forme :

$$C(v) = \frac{1}{\lambda} \sum_{u \neq v} A(v,u) C(u) \tag{13.5}$$

qui peut s'écrire $\lambda C = AC$.
Pour résoudre cette équation, BONACICH 1987 montre que le vecteur de centralité spectrale correspond en fait au vecteur propre dominant (ou principal) de la matrice d'adjacence.

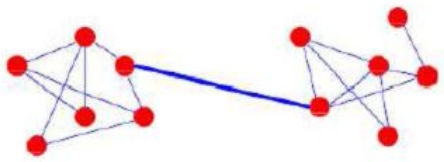


FIGURE 13.9 – Une forte centralité d'intermédiarité (trait foncé)

On peut illustrer ces concepts et montrer dans quelle mesure ils diffèrent en utilisant une des bases de données les plus classiques, celle de Zachary (ZACHARY 1977) sur le réseau social constitué par les membres d'un club de karaté universitaire (figure 13.10). Le package *igraph* du logiciel R permet de représenter le graphe et de calculer les indicateurs précédents.

```
# Centralite de degre
d<- degree(kar)
# Centralite de proximite
cp<- closeness(kar)
# Centralite d'intermediarite
```

```
ci<- betweenness(kar)
# Centralite de vecteur propre
ce<- graph. eigen(kar)[c("values", "vectors")]
```

```
kar <- read.graph("karate.gml",format="gml")
plot(kar)
```

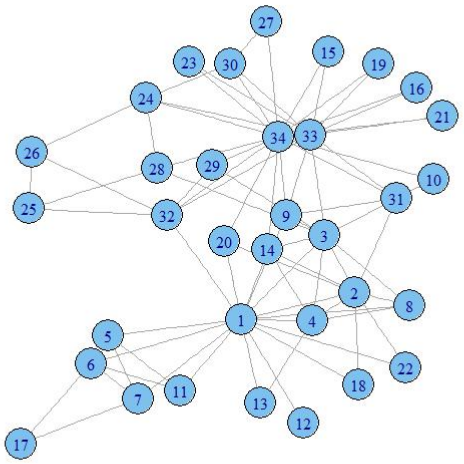


FIGURE 13.10 – Réseau de Zachary
Note : liens d'amitié entre 34 membres d'un club de karaté dans une université américaine

Le tableau ci-dessous montre le classement des individus du réseau présenté en figure 13.10 selon les différents critères de centralité. Le classement est assez concordant pour les premiers du classement. Six individus partagent les cinq premières places de chaque indicateur. L'individu 1 est toujours dans les deux premières positions, notamment pour la proximité et l'intermédiarité. Il doit cette position au fait qu'il a un grand nombre de liens (centralité de degré élevé) et qu'il est l'intermédiaire obligé pour un petit groupe d'individus (centralité d'intermédiarité forte) qui sont eux-mêmes peu liés aux autres. Ainsi il est proche de tous les autres membres du club, soit une forte centralité de proximité. La centralité de vecteur propre résume ces notions.

Classement pour chaque indicateur	Degré	Proximité	Intermédiarité	Vecteur propre
Premier	34	1	1	34
Deuxième	1	3	34	1
Troisième	32	34	33	3
Quatrième	3	32	3	33
Cinquième	2	33	32	2

13.2.2 Les méthodes de partitionnement

Si l'on revient à nos problèmes de réseaux de villes, on va être confronté à la détermination de communautés. Dans le premier chapitre, on a vu que les réseaux de villes combinent souvent

des aspects "petit-monde", avec de forts liens en intra, et des aspects invariants d'échelle, avec des sous-groupes assez fortement différenciés. On s'appuiera largement dans cette partie sur les synthèses réalisées par NEWMAN 2006 et FORTUNATO 2010, ainsi que sur les thèses francophones de PONS 2007 et SEIFI 2012.

Définition et qualité d'une partition

Le premier problème du partitionnement de graphes est celui de la définition d'une communauté. Aucune définition n'est universellement acceptée. Ce qui unifie les approches, sans déboucher sur une définition précise, c'est qu'il doit y avoir plus de liens au sein de la communauté que de liens vers le reste du graphe. Cela ne peut se produire que si les graphes sont peu denses, clairsemés (*sparse*), et si le nombre de liens reste du même ordre de grandeur que celui de sommets.

Les graphes associés aux réseaux sociaux, ou certains graphes décrivant des structures biologiques, atteignent de très grandes tailles, contrairement à ceux que l'on a présentés jusqu'à présent. Le partitionnement de ces graphes en communautés nécessite des algorithmes très performants. Leur nombre est croissant. Ils utilisent des méthodes souvent issues de la physique (méthodes "gloutonnes", *spinglass*).

Comme en classification, on sera confronté au problème d'optimisation du nombre de communautés, à celui de hiérarchie et à celui d'emboîtement.

Les communautés peuvent être appréhendées d'un point de vue local, c'est-à-dire en faisant le plus possible abstraction du graphe perçu comme un tout. Dans cette perspective, on privilégie les indicateurs qui mesurent la cohésion interne, qu'on pourrait traduire dans le langage des réseaux sociaux par le fait que tout le monde est ami de tout le monde. Dans ces communautés, on doit voir apparaître beaucoup de **cliques** (sous-graphes maximaux complets comprenant au moins trois sommets). On s'intéresse aussi de ce point de vue à la densité des liens au sein de la communauté et à celle des liens qui la relient au reste du graphe.

Elles peuvent aussi être définies en considérant le graphe comme un tout. Une des idées essentielles est de comparer la structure d'un graphe présentant des communautés à celle d'un graphe aléatoire. Ces graphes, souvent qualifiés de graphes d'Erdos-Renyi ont été les premiers étudiés. Si l'on cherche encore une fois des analogies avec les méthodes statistiques, on cherche un *modèle nul* auquel comparer notre graphe réel. Ce modèle nul doit être un graphe aléatoire, bien sûr, mais qui respecte, pour qu'il soit comparable, un certain nombre de contraintes. La version la plus utilisée est celle qui a été proposée par NEWMAN et al. 2004. Elle consiste en une version "randomisée" du graphe original, c'est-à-dire où les liens sont modifiés de façon aléatoire, sous la contrainte que le degré attendu de chaque sommet corresponde à celui du graphe original. Cette approche a permis à ces auteurs de proposer une des notions les plus fécondes en théorie du partitionnement, celle de modularité.

La modularité permet de justifier la pertinence des sous-graphes obtenus après un partitionnement. L'hypothèse forte de la modularité est la comparaison avec un graphe aléatoire, ce qui sous-entend qu'un graphe ayant une structure complètement aléatoire doit avoir une modularité proche de 0. Cette comparaison permet donc de mettre en évidence des relations plus denses que la moyenne, soit une structure communautaire, ou à l'inverse si les relations sont moins denses, des structures isolées.

Définition 13.2.3 — La modularité. C'est une mesure de la qualité d'un partitionnement de graphe. Si l'on considère $\mathbf{P}$ la partition en p clusters du graphe $G = \{V, E\}$, alors : $\mathbf{P} = \{c_1, ..., c_n, ...c_p\}$

La modularité peut être introduite de façon assez simple de la façon suivante, en se référant

à l'idée de Newman.

$$Q(P) = \sum_i (e_{c_i} - a_{c_i}^2) \tag{13.6}$$

avec e_{c_i} la part des liens d'un cluster c_i sur le total, a_{c_i} la probabilité qu'un sommet se trouve dans le cluster c_i et donc $a_{c_i}^2$ la probabilité que les deux sommets d'un lien se trouvent dans le même cluster c_i.

Cette expression générale est transformée dans la première forme usuelle de présentation de la modularité. On montre (FORTUNATO 2010) que la modularité peut s'écrire sous la forme :

$$Q(P) = \frac{1}{2m} \sum_{i,j \in V} \left(A_{ij} - \frac{d_i d_j}{2m} \right) \delta(c_i, c_j) \tag{13.7}$$

avec

— m le nombre d'arêtes du graphe ;
— A la matrice d'adjacence du graphe ;
— A_{ij} le poids des liens entre les sommets i et j ;
— d_i la somme des degrés de i avec $d_i = \sum_i A_{ij}$;
— $a_{c_i}^2 = \sum_j \frac{d_i d_j}{4m^2}$;
— $\delta(c_i, c_j)$ une fonction de Kronecker qui vaut 1 si les deux sommets appartiennent à la même communauté et 0 sinon.

On peut montrer qu'une façon alternative d'écrire cette expression est la suivante :

$$Q(P) = \frac{1}{2m} \sum_{k=1}^{p} \sum_{i,j \in V} \left(A_{ij} - \frac{d_i d_j}{2m} \right) = \sum_{k=1}^{p} \left[\frac{l_k}{m} - (\frac{d_k}{2m})^2 \right] \tag{13.8}$$

où l_k désignant le nombre de liens joignant les sommets de la communauté k et d_k la somme des degrés de la communauté k.

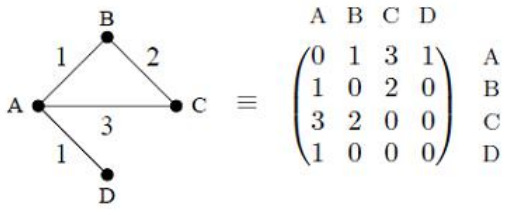
Le terme $A_{ij} - \frac{d_i d_j}{2m}$ correspond à la différence de liens entre notre graphe et un graphe aléatoire dont la contrainte est la conservation des degrés de sommets.

Les définitions de la modularité ont d'abord été développées dans le contexte de graphes non valués. Elles ont été étendues aux graphes valués. La valeur de A_{ij} correspond au lien entre les sommets, qui dans un graphe non valué, vaut 1 si les sommets sont liés et 0 sinon, et dans un graphe valué vaut la valeur du flux s'il y a un lien et 0 sinon. On trouve dans NEWMAN 2004 une façon très simple de passer des graphes non valués aux graphes valués, en introduisant ce qu'il appelait des multigraphes (figure 13.11).

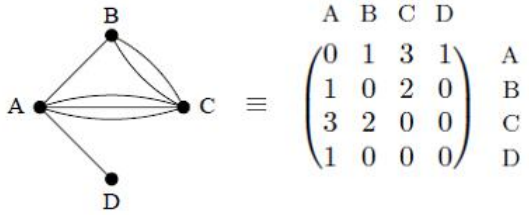
Cette représentation permet de généraliser aux graphes pondérés les résultats présentés précédemment. Les A_{ij} correspondent aux poids associés aux liens ou de façon équivalente au nombre de liens du multigraphe. M est le nombre de liens du multigraphe, ou la somme des pondérations.

La modularité est un des concepts les plus puissants de la théorie des partitions de graphe, et malgré les critiques émises à son encontre, le plus utilisé. Il est utilisé comme fondement de certaines méthodes, et comme mesure de la qualité de partitions produites par d'autres méthodes. On l'utilisera à plusieurs reprises dans les exemples que l'on donnera.

Les travaux de Guimaras, Reichart et Bornholdt, mis en avant par Fortunato (FORTUNATO 2010) se penchent sur le problème de "résolution". Si le nombre de liens dans le graphe devient très grand et que le nombre de liens attendu (voir formule de la modularité) est inférieur à 1, un seul lien entre les deux groupes suffit à entraîner leur fusion.



(a) Graphe valué



(b) Représentation multigraphe

FIGURE 13.11 – Passage au graphe valué, les multigraphes

Panorama général des méthodes de partition

Une fois défini le schéma général d'une partition, il reste à la réaliser en pratique. En pratique dans ce cas implique de trouver des façons de faire, des algorithmes donc, qui permettent d'une part de résoudre le problème, et ensuite de le résoudre dans un temps acceptable. Les graphes des réseaux de villes sont déjà conséquents mais restent très petits si on les compare à ceux des réseaux sociaux ou même à ceux qui sont utilisés dans l'étude des protéines ou du génome. La complexité des algorithmes (problèmes NP-difficiles ou NP-complets) est présentée dans FORTUNATO 2010. On cherche souvent à mesurer la complexité des algorithmes en les notant $O(n^2m^2)$ avec n le nombre de liens, m le nombre d'arêtes. On va retrouver dans les méthodes des questions bien connues en analyse des données : combien de classes ? doit-on les déterminer au préalable ? doit-on appliquer des méthodes ascendantes ou descendantes ? comment déterminer des critères d'arrêt ? On se limitera ici à la présentation de quelques familles de méthodes testées dans le cadre des travaux du pôle "Analyses Territoriales" de l'Insee sur les réseaux de villes (voir section 13.3) et en se centrant sur celles qui sont implémentées dans le logiciel R. Les méthodes sont en pleine expansion et font l'objet de controverses au sein des spécialistes. Beaucoup de celles qui sont présentées ici sont issues des travaux de Mark Newman, introducteur entre autres de la notion de modularité présentée dans le paragraphe précédent. La complexité algorithmique des questions a fait que beaucoup de travaux initiaux ont porté sur la bipartition des graphes (KERNIGHAN et al. 1970). D'autres méthodes s'inspiraient aussi de ce qui était fait en analyse des données (dendrogrammes de classification, méthodes de type *k-means*). Ces méthodes reposent sur des propriétés des graphes, ou sur le traitement de la matrice d'adjacence.

Méthodes classiques

On ne présentera que quelques unes des méthodes classiques :

Les méthodes fondées sur la bissection de graphes

Ces méthodes (figure 13.12) sont assez simples à présenter. L'idée est de chercher la ligne qui partage le graphe en coupant le moins de liens (*cut size*).

Dans leurs versions les plus simples, ces méthodes risquent cependant de ne faire apparaître que des solutions triviales (un sommet isolé). Des méthodes plus élaborées de bissection reposent

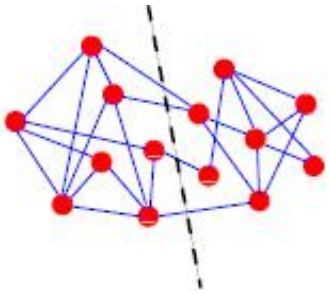


FIGURE 13.12 – Bipartition de graphe

sur des méthodes spectrales (propriétés du spectre de la matrice laplacienne) que l'on présentera plus loin.

Les méthodes hiérarchiques

Ces méthodes (figure 13.13) reposent sur des mesures de similarité entre les sommets. Lorsqu'on a calculé cette similarité pour chaque paire de sommets (matrice de similarité), on peut construire par exemple un dendrogramme par des méthodes assez classiques.

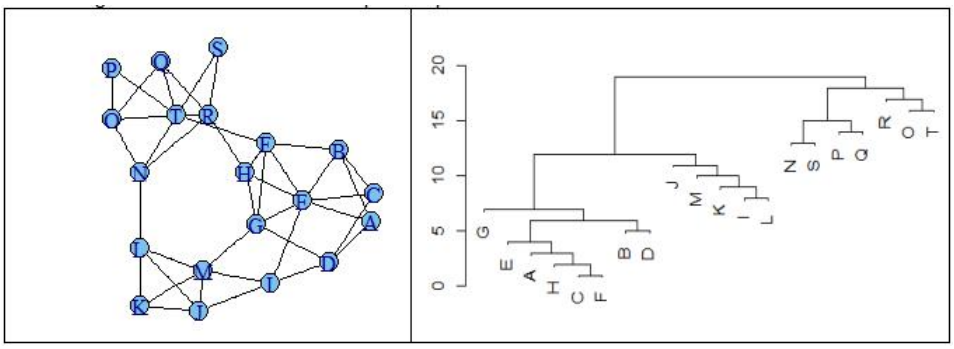


FIGURE 13.13 – Méthodes hiérarchiques de partitionnement

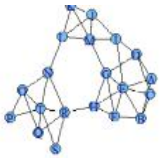
Les méthodes de *clustering*

Ces méthodes sont bien connues en analyse des données. Dans ces méthodes, le nombre de classes est prédéterminé. On définit une distance entre couples de points, d'autant plus grande que les sommets sont dissemblables. On va chercher à minimiser une fonction de coût basée sur les points et les centroïdes. Dans le minimum *k-means clustering*, par exemple, la fonction de coût est la plus grande distance entre deux points de la classe. On cherche à trouver la partition qui rende minimale la plus grande des k classes (recherche de classes compactes). La méthode de MacQueen repose elle sur la minimisation du total des distances intra-classes.

La méthode divisive

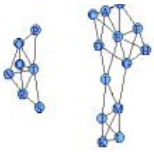
Cette méthode est une des plus intuitives à présenter. Elle repose sur le concept présenté en 13.2.1 de centralité d'intermédiarité, avec un schéma qui expose assez bien dans un cas simple cette idée. Lorsque beaucoup de géodésiques allant d'un point quelconque du graphe à un autre passent par un sommet ou par un lien, la suppression de ceux-ci est plus à même de faire apparaître des communautés. Dans l'exemple présenté plus haut, c'est le lien entre les sommets T et F qui a la plus forte centralité d'intermédiarité. Si on supprime ce lien, c'est le lien RH qui a alors la plus forte centralité, puis le lien NL. Cette démarche est représentée en figure 13.14.

Après la suppression de ces trois liens, le graphe n'est plus connexe et une communauté apparaît. Le processus peut se poursuivre. Une commande de R produit le résultat final.



(a) Suppression du lien T-F

(b) Suppression du lien R-H



(c) Suppression du lien N-L

FIGURE 13.14 – Partitionnement du graphe de de la figure 13.13 avec la méthode divisive

```
karate <- read.graph("karate.gml",format="gml")
plot(karate,vertex.size=2)
betkar<-  edge.betweenness.community(karate)
plot(betkar,karate)
```

Le résultat sur ce graphe très simple est assez trivial et on peut voir ce qu'il produit sur un graphe encore lisible mais plus complexe comme celui du club de karaté. La méthode divisive la plus connue est celle de NEWMAN et al. 2004. Elle confirme d'ailleurs l'attrait des physiciens pour l'étude des graphes. L'algorithme illustré précédemment est le suivant :

1. calcul de la centralité d'intermédiarité pour tous les liens ;
2. suppression du lien ayant la plus forte centralité ;
3. re-calcul de la centralité ;
4. itération du cycle à l'étape 2.

Ce processus itératif peut se poursuivre jusqu'à l'isolement de tous les sommets et produire ainsi une hiérarchie de partitions emboîtées. Le choix de la partition peut se faire avec le critère de modularité. Cet algorithme nécessite à chaque étape le calcul des centralités d'intermédiarité et sa complexité est en $O(m^2n)$, ce qui le rend inexploitable sur de très grands graphes.

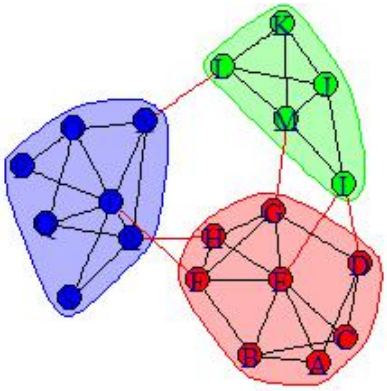
D'autres algorithmes divisifs ont été proposés. FORTUNATO 2010 a proposé un algorithme qui utilise la centralité d'information de lien définie comme étant la diminution relative de l'efficacité du réseau lorsque l'on retire ce lien du graphe. Cet algorithme est plus performant, mais de complexité plus grande que celui de Girvan-Newman. Ce dernier reste donc très utilisé, notamment à titre de comparaison des communautés détectées.

Les méthodes agglomératives fondées sur la modularité

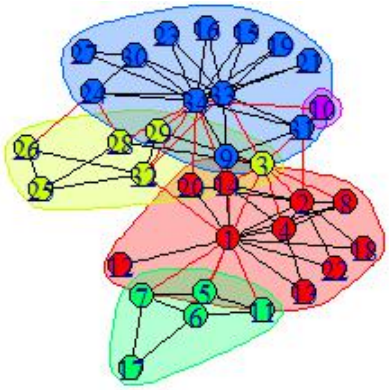
Cette famille de méthodes est très riche et très importante. Au contraire de la précédente, on part de l'ensemble des sommets, que l'on va progressivement agréger entre eux.

La méthode "optimale"

Elle repose sur l'exploration de toutes les communautés possibles et sur la maximisation de la modularité. On peut trouver dans FORTUNATO 2010 une valeur approchée du nombre de ces



(a) Graphe de démonstration 3 communautés



(b) Club de Karaté 5 communautés

FIGURE 13.15 – Résultat de la méthode divisive sur un graphe et sur le club de karaté

partitions, nombre qui explose avec la taille du graphe et la rend inexploitable, même pour des graphes de taille moyenne. Le calcul des communautés dans cette optique utilise une méthode issue de la physique appelée "recuit simulé" (succession d'allers/retours à l'état d'équilibre) souvent utilisée dans les problèmes d'optimisation. Elle est implémentée en R dans le package *igraph* par la commande `optimal.community`.

La méthode de Clauset et Newman - méthode aggrégative

C'est un algorithme qualifié de "glouton" qui permet la constitution d'une partition à partir d'un critère de modularité. Il a d'abord été proposé par Newman en 2003 puis par Clauset, Newman et Moore dans une deuxième version. Il utilise la modularité sous la forme suivante : $Q = \sum_i (e_i - a_i^2)$. On définit une grandeur notée ΔQ_{ij} correspondant à la variation de modularité lorsqu'on fait un lien entre la communauté i et la communauté j. Le détail de l'algorithme, avec les indications liées au stockage de l'information peuvent être trouvées dans CLAUSET et al. 2004. Le schéma général est le suivant :

1. on part de n communautés (chaque sommet étant une communauté) ;
2. on calcule ΔQ_{ij} pour toutes les paires ;
3. on fusionne les paires qui accroissent le plus la modularité ;
4. on répète les phases 2 et 3 jusqu'à ce qu'on obtienne une seule communauté ;
5. on coupe le dendrogramme à la valeur correspondant à la plus forte modularité.

Dans cet exemple très simple (figure 13.16), on peut voir que la modularité Q augmente jusqu'à l'étape 10 où les trois communautés assez visibles sont identifiées. À l'étape 11, deux des communautés fusionnent et la modularité diminue, celle-ci devenant nulle lorsque les trois communautés sont regroupées. Le résultat est donc un partitionnement en trois communautés avec une modularité de $0,485$. Cet algorithme est implémenté en R dans le package *igraph* par la fonction `fastgreedy.community`. La caractéristique de cet algorithme est sa grande vitesse d'exécution qui lui permet de traiter de grands graphes. L'algorithme est de complexité $O(mn)$.

Les méthodes spectrales

Newman a proposé une version spectrale du partitionnement fondée sur la modularité. Dans cette version, on introduit une matrice qui fait apparaître l'expression de la modularité : $B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$. Dans le cas initial d'une bipartition, généralisée ultérieurement, Newman introduisait un vecteur s valant $+1$ si le sommet appartenait au premier groupe, -1 au second. Il montre que la maximisation de la modularité en fonction du vecteur s se ramène à un problème que l'on peut formaliser par : $B_s = \lambda D_s$ dans lequel λ est un multiplicateur de Lagrange, et D une matrice diagonale contenant les degrés des sommets. Lorsqu'on résout ce problème matriciel, compte tenu de la structure de la matrice sur laquelle on travaille, on obtient une solution triviale avec une valeur propre égale à 0 et un vecteur composé de 1, soit le regroupement de tous les sommets dans une seule communauté. Pour effectuer la partition, on utilise le vecteur propre associé à la plus grande valeur propre (NEWMAN 2006). On trouve dans le package *igraph* la fonction `leading.eigenvector.community` qui met en œuvre cette méthode.

Algorithme de Louvain

En 2008, trois chercheurs de l'université de Louvain ont proposé une autre méthode "gloutonne", plus rapide que la majorité des autres approches. Sa particularité est de se fonder sur une approche locale de la modularité. Dans une première phase, une communauté différente est attribuée à chaque sommet. On s'intéresse ensuite aux voisins de chaque sommet i, et on calcule le gain de modularité en retirant le sommet i et en le plaçant dans la communauté j. On recherche un gain positif et maximum pour déplacer i. On effectue cette opération de façon séquentielle jusqu'à ce qu'aucune amélioration ne soit possible. La deuxième phase de l'algorithme consiste à construire un nouveau réseau dont les sommets sont les communautés repérées dans la première phase, les

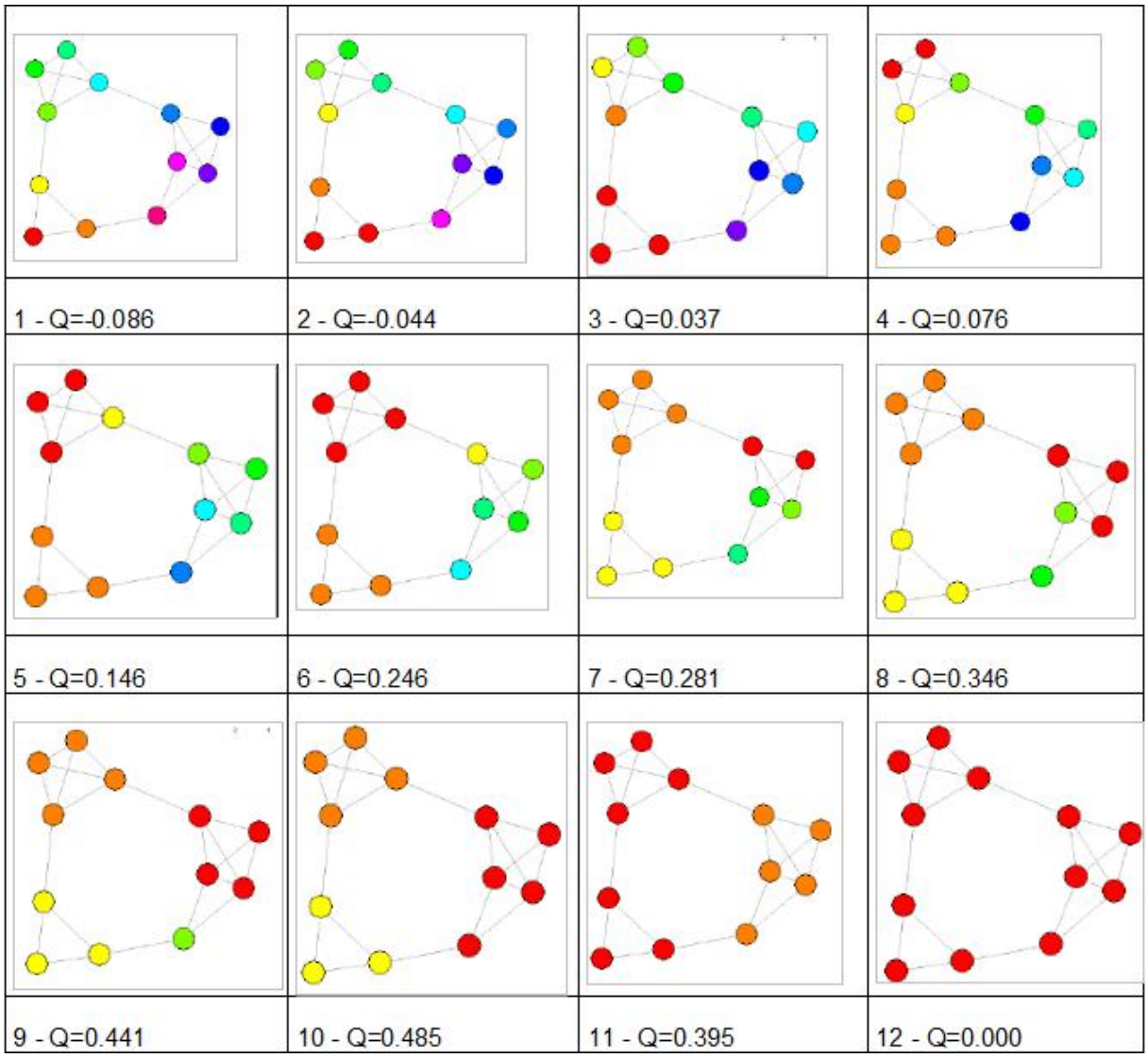


FIGURE 13.16 – Les 12 étapes du partitionnement d'un graphe à 12 sommets avec la méthode aggrégative

poids des liens entre les communautés étant déterminés par la somme des poids des liens des sommets du graphe initial. Une fois cette deuxième phase terminée, on ré-applique l'algorithme à ce nouveau réseau pondéré. Une combinaison des deux phases est une "passe", et ces passes sont itérées jusqu'à ce qu'un maximum de modularité soit atteint. On trouve dans le package *igraph* la fonction `multilevel.community` qui met en œuvre cette méthode. Elle est souvent présentée, notamment dans des articles récents de Newman comme la plus performante en temps et en qualité de partitionnement (NEWMAN 2016).

Autres méthodes

Marches aléatoires (*random walk*)

L'algorithme `walktrap.community` vise au final, comme tous les autres, à produire des distances entre les sommets du graphe. L'idée est d'aboutir à cette distance en se fondant sur l'idée de marche aléatoire. Le temps est discrétisé. À chaque instant, un marcheur se déplace aléatoirement d'un sommet vers un sommet choisi parmi ses voisins. La suite des sommets visités est alors une marche aléatoire. La probabilité d'aller du sommet i au sommet j est :

$$P_{ij} = \frac{A_{ji}}{k_i}. \tag{13.9}$$

On a ainsi la matrice de transition de la chaîne de Markov correspondante, et on peut calculer la probabilité de passer du sommet i au sommet j en un temps t, $P_{ij}(t)$. Lors d'une marche aléatoire suffisamment longue dans un graphe, la probabilité de se trouver sur un sommet donné est directement (et uniquement) proportionnelle au degré de ce sommet. La probabilité d'aller de i à j et celle d'aller de j à i par une marche aléatoire de longueur fixée ont un rapport de proportionnalité qui ne dépend que des degrés des sommets de départ et d'arrivée :

$$k_i P_{ij}(t) = k_j P_{ji}(t). \tag{13.10}$$

La façon de comparer deux sommets i et j doit s'appuyer sur les constatations suivantes :

— si deux sommets i et j sont dans une même communauté, la probabilité $P_{ij}(t)$ est certainement élevée. En revanche si $P_{ij}(t)$ est élevée, il n'est pas toujours garanti que i et j soient dans la même communauté ;

— la probabilité $P_{ij}(t)$ est influencée par le degré k_j du sommet d'arrivée : les marches aléatoires ont plus de chances de passer par les sommets de fort degré (dans le cas limite d'une marche aléatoire infinie, cette probabilité est proportionnelle au degré) ;

— les sommets d'une même communauté ont tendance à voir les sommets éloignés de la même façon, ainsi si i et j sont dans la même communauté et k dans une autre communauté ; il y a de fortes chances que $P_{ik}(t) = P_{jk}(t)$. On définit ainsi une distance, qui doit être plus faible lorsque les deux sommets appartiennent à la même communauté :

$$\sqrt{\sum_{k=1}^{n} \frac{(P_{ik}(t) - P_{jk}(t))^2}{k_k}}. \tag{13.11}$$

Dans cette méthode, le choix de t est très important. Si t est trop petit, les communautés sont minuscules. S'il est trop grand, les probabilités tendent vers la même valeur. Une fois déterminée la matrice de distances, l'algorithme est assez classique : on part de n communautés et on agrège ensuite. On obtient un arbre et on utilise la modularité pour trouver la partition adaptée. On trouvera les détails dans PONS 2007.

Dans l'exemple de la figure 13.17, on a, jusqu'à $t = 3$, représenté graphiquement la matrice de probabilité qui sera utilisée pour faire le partitionnement (par analyse spectrale). On trouve dans le package *igraph* la fonction `walktrap.community` qui met en œuvre cette méthode.

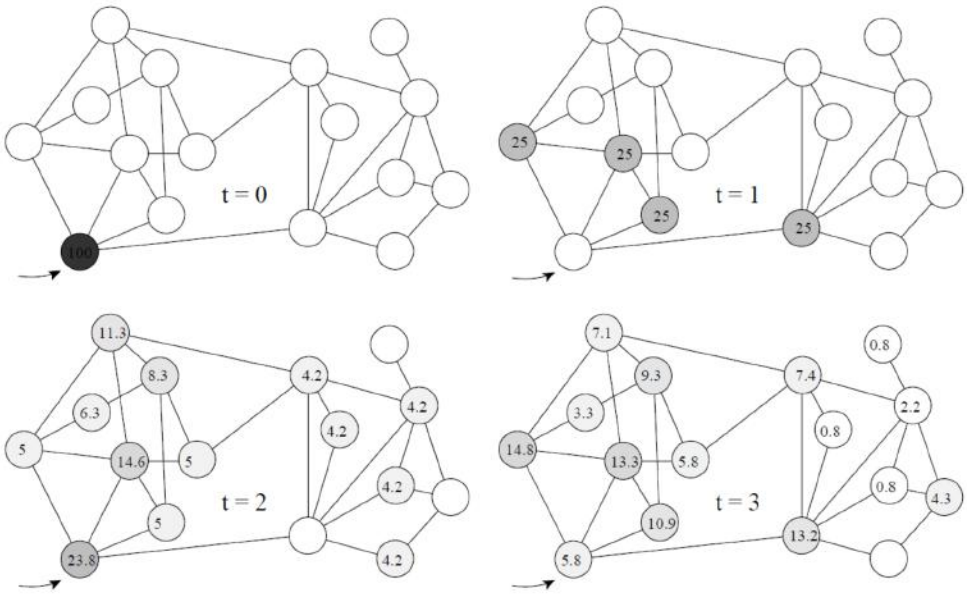


FIGURE 13.17 – Illustration de la marche aléatoire sur un graphe
Source : *d'après* PONS *2007*

Verres de spin

Avec cette méthode, on s'éloigne des méthodes usuelles. Elle s'inspire des verres de spin, qui sont des alliages correspondant à des impuretés, un spin étant associé à chaque impureté. Le couplage entre les différents spins peut être plus ou moins intense. Cette méthode est utilisée en physique théorique. Les paires de spins sont associées dans un graphe. On définit un **graphe hamiltonien** (graphe possédant au moins un cycle passant par tous les sommets une fois au plus) et une distribution de probabilité des couplages. REICHARDT et al. 2006 ont utilisé cette approche. Chaque sommet est caractérisé par un spin prenant q valeurs possibles, et les communautés correspondent aux valeurs de sommets ayant des valeurs de spins égales. On définit l'énergie du système par un hamiltonien faisant intervenir la matrice d'adjacence du graphe. La minimisation de cette expression se fait par recuit simulé, comme pour la méthode "optimale" présentée précédemment. On trouve dans le package *igraph* la fonction `spinglass.community` qui met en œuvre cette méthode.

Références - Chapitre 13

BARABÁSI, Albert-László et Réka ALBERT (1999). « Emergence of scaling in random networks ». *Science* 286.5439, p. 509–512.

BATTISTON, Federico, Vincenzo NICOSIA et Vito LATORA (2014). « Structural measures for multiplex networks ». *Physical Review E* 89.3, p. 032804.

BEAUGUITTE, Laurent et César DUCRUET (2011). « Scale-free and small-world networks in geographical research : A critical examination ». *17th European Colloquium on Theoretical and Quantitative Geography*, p. 663–671.

BONACICH, Phillip (1987). « Power and centrality : A family of measures ». *American journal of sociology* 92.5, p. 1170–1182.

CHRISTALLER, Walter (2005). « Les lieux centraux en Allemagne du Sud Une recherche économico-géographique sur la régularité de la diffusion et du développement de l'habitat urbain ». *Cybergeo : European Journal of Geography*.

CLAUSET, Aaron, Mark EJ NEWMAN et Cristopher MOORE (2004). « Finding community structure in very large networks ». *Physical review E* 70.6, p. 066111.

FORTUNATO, Santo (2010). « Community detection in graphs ». *Physics reports* 486.3, p. 75–174.

KARINTHY, Frigyes (1929). « Chain-links ». *Everything is the Other Way*, p. 25.

KERNIGHAN, Brian W et Shen LIN (1970). « An efficient heuristic procedure for partitioning graphs ». *The Bell system technical journal* 49.2, p. 291–307.

NEWMAN, Mark EJ (2004). « Analysis of weighted networks ». *Physical review E* 70.5, p. 056131.

— (2006). « Modularity and community structure in networks ». *Proceedings of the national academy of sciences* 103.23, p. 8577–8582.

NEWMAN, Mark EJ et Michelle GIRVAN (2004). « Finding and evaluating community structure in networks ». *Physical review E* 69.2, p. 026113.

NEWMAN, Mark, Albert-Laszlo BARABASI et Duncan J WATTS (2011). *The structure and dynamics of networks*. Princeton University Press.

NEWMAN, MEJ (2016). « Community detection in networks : Modularity optimization and maximum likelihood are equivalent ». *arXiv preprint arXiv :1606.02319*.

PONS, Pascal (2007). « Détection de communautés dans les grands graphes de terrain ». Thèse de doct. Paris 7.

REICHARDT, Jörg et Stefan BORNHOLDT (2006). « Statistical mechanics of community detection ». *Physical Review E* 74.1, p. 016110.

ROZENBLAT, Céline et Guy MELANÇON (2013). *Methods for multilevel analysis and visualisation of geographical networks*. Springer.

SEIFI, Massoud (2012). « Cœurs stables de communautés dans les graphes de terrain ». Thèse de doct.

WATTS, Duncan J et Steven H STROGATZ (1998). « Collective dynamics of'small-world'networks ». *nature* 393.6684, p. 440.

WILSON, Alan Geoffrey (1974). *Urban and regional models in geography and planning*. John Wiley & Sons Inc.

ZACHARY, Wayne W (1977). « An information flow model for conflict and fission in small groups ». *Journal of anthropological research* 33.4, p. 452–473.

14. Confidentialité des données spatiales

MAËL-LUC BURON, MAËLLE FONTAINE
Insee

Résumé

La profusion récente de sources de données géolocalisées, souvent diffusées sous forme de données carroyées, offre de nombreux champs aux économistes, démographes ou sociologues. Toutefois, cette profusion entraîne un risque élevé de divulguer de l'information confidentielle. En effet, le nombre de variables nécessaires pour identifier de manière unique une personne diminue considérablement une fois que l'on connaît sa position géographique. Le risque de divulgation est encore plus élevé dans les zones où la densité de population est faible.

Traditionnellement, les méthodes de gestion de la confidentialité de données statistiques ne tiennent pas compte de l'information spatiale présente dans les données. Ce chapitre vise à pallier en partie ce manque, en présentant des méthodes qui gèrent la confidentialité des données tout en préservant leur utilité en termes de corrélations spatiales. Les méthodes prétabulées semblent davantage correspondre au but recherché, car elles ciblent les observations les plus exposées au risque, celui-ci étant déterminé en fonction du contexte local. Cependant, appliquer uniquement des méthodes prétabulées ne suffit pas à atteindre un niveau de protection suffisant, et des méthodes post-tabulées peuvent être mises en œuvre dans un deuxième temps pour garantir le secret statistique.

Dans ce chapitre, nous présentons la littérature existante, en particulier une méthode spécifique, appelée *targeted record swapping*, mise en avant dans le programme d'Eurostat "Protection harmonisée des données de recensement au sein du système statistique européen". Le principe de

cette méthode est de détecter les observations les plus exposées au risque de divulgation et de les échanger avec d'autres observations "proches". Ainsi, les observations qui présentent des caractéristiques rares sont toujours présentes dans les données, mais pas avec leur localisation géographique réelle, ce qui empêche l'intrus de les ré-identifier avec certitude. Nous avons testé cette méthode sur des données fiscales françaises à l'échelle d'une petite région, et pour plusieurs jeux de paramètres. Nous obtenons une très faible déformation des corrélations spatiales pour les variables prises en compte dans les paramètres de la méthode ou fortement corrélées avec celles-ci.

R La lecture préalable du chapitre 1 : "Codifier la structure de voisinage" et du chapitre 3 : "Indices d'autocorrélation spatiale" est recommandée.

Introduction

Disposer de données spatiales permet au statisticien de révéler et d'expliquer des phénomènes sous-jacents. Les chapitres précédents dressent un panorama des outils d'analyse possibles pour tirer profit de l'information spatiale.

Aujourd'hui, de plus en plus de données sont géolocalisées, ce qui rend possible de diffuser de l'information statistique à un niveau géographique fin. De nombreux sujets d'étude émergent donc pour les analystes. Cependant, la contrepartie de cette profusion de données spatiales est un enjeu crucial lié à leur confidentialité. En effet, le nombre de caractéristiques nécessaires pour identifier de manière unique un individu statistique diminue avec la taille de la maille géographique de diffusion, et ce d'autant plus dans le contexte actuel de prolifération d'outils libres de visualisation géographique. En outre, lorsque la densité de population est faible dans une zone donnée, le risque de divulgation augmente, car la probabilité de trouver un individu similaire dans le voisinage est faible.

Dans ce chapitre, deux grands principes de la diffusion de statistiques publiques s'opposent (VANWEY et al. 2005). D'un côté, les Instituts Nationaux de Statistique (INS) ont la vocation de diffuser des données avec le plus grand niveau d'utilité possible, et de l'autre, ils doivent respecter de fortes contraintes de confidentialité des enquêtés. Dans le cas des données spatiales, garantir la confidentialité est une tâche particulièrement difficile, car les réglementations européennes et nationales interdisent aux INS de diffuser toute donnée susceptible de permettre à un intrus de trouver, directement ou indirectement, l'identité du ménage ou de l'entreprise enquêté(e). Au sens strict, cela signifierait, dans la plupart des cas, qu'ils ne peuvent rien publier du tout, car le risque zéro n'existe pas dès lors que l'on publie des données. Aussi l'objectif est plutôt de le ramener à un niveau faible, jugé acceptable. En d'autres termes, la stratégie de protection des données peut être perçue comme un compromis à trouver entre une minimisation du risque de divulgation et une maximisation de l'utilité des données.

Ce chapitre n'est pas écrit du point de vue de l'utilisateur des données, mais de celui de l'expert en confidentialité statistique, dont le rôle est de garantir que les données diffusées respectent le secret statistique. En général, cet expert dispose d'un fichier de données individuelles (microdonnées) et doit diffuser des données à un niveau géographique fin (région, carreau, etc.), mais il lui est interdit de publier une statistique si elle concerne moins d'un certain nombre-seuil d'observations. Il met donc en œuvre une procédure de confidentialisation appelée dans la suite méthode SDC (pour *Statistical Disclosure Control*). Dans ce chapitre, une observation peut se rapporter à un ménage, à un individu ou à une entreprise. Nous supposons que les microdonnées sont exhaustives : sans adaptations supplémentaires, les méthodes présentées ne doivent pas être appliquées aux données d'enquêtes.

La section 14.1 présente le risque de divulgation : comment celui-ci peut être défini dans le cas de données spatiales, et quelles recommandations peuvent être formulées pour détecter

les observations les plus risquées. Les méthodes standard utilisées pour gérer la confidentialité statistique ont fait l'objet d'un manuel publié par Eurostat en 2007 puis dans une seconde version en 2010 (HUNDEPOOL et al. 2010), mais les données spatiales requièrent des adaptations de ces méthodes. La section 14.2 donne un panorama de différentes méthodes SDC adaptées aux données spatiales, et des considérations sur les analyses risque-utilité. La section 14.3 présente les résultats d'une méthode prétabulée testée à l'échelle d'une région française, dans le contexte de diffusion de données carroyées. Pour ces tests, les problèmes de différenciation avec les zonages administratifs ne sont pas examinés, mais la section 14.4 est spécifiquement consacrée à ce sujet.

14.1 Comment évaluer le risque de divulgation spatiale ?

14.1.1 Définition générale du risque de divulgation

Afin de garantir à chacun la protection de ses données personnelles, des règlements européens ont été rédigés pour imposer le secret statistique [1]. Ainsi, d'après l'article 20 du chapitre V du Règlement n°223/2009 du Parlement Européen relatif aux statistiques européennes : "Dans leurs domaines de compétence respectifs, les INS et autres autorités nationales ainsi que la Commission (Eurostat) prennent toutes les mesures réglementaires, administratives, techniques et organisationnelles nécessaires pour assurer la protection physique et logique des données confidentielles (contrôle de la confidentialité statistique). ". Les pays appliquent également leur propre réglementation. Les contraintes de confidentialité prennent en général la forme de seuils idoines : aucune information ne peut être communiquée si elle concerne moins d'un certain nombre d'observations. Le choix de ces seuils dépend de différents éléments : densité de population, aversion au risque, degré de sensibilité des variables diffusées, nature des utilisateurs. Parfois, des recommandations explicites sont disponibles pour vérifier si le fichier de données respecte les règles de confidentialité (ONS 2006, INSEE 2010).

On parle de **divulgation** lorsqu'un intrus (également appelé *data snooper* dans certains articles) utilise des données diffusées pour obtenir des renseignements inconnus auparavant. Cet intrus n'agit pas de façon illégale et ne tente pas de casser un système de sécurité : il ne mobilise que les données mises à sa disposition. On distingue en général différents scénarios de divulgation (DUNCAN et al. 1986, LAMBERT 1993, CLIFTON et al. 2012 [2], BERGEAT 2016) :

- **la divulgation d'identité** se produit lorsqu'un identifiant direct d'un individu statistique (entreprise, ménage ou individu) peut être retrouvé grâce à des données diffusées (par exemple, il peut être facile d'identifier l'entreprise qui fait le plus gros chiffre d'affaires dans un secteur donné) ;
- **la divulgation d'attributs** survient lorsque l'intrus peut accéder à de l'information sensible (variables appelées "**quasi-identifiants**" dans la suite) d'un individu. La divulgation d'identité implique toujours une divulgation d'attributs, mais le contraire n'est pas vrai : par exemple, si l'intrus connaît un habitant d'une zone, et si les données diffusées indiquent que tous les habitants de cette zone partagent une caractéristique commune, l'intrus peut déduire que l'individu présente cette caractéristique, même s'il ne déduit pas les autres attributs de cet individu ;
- **la divulgation inférentielle** se produit lorsqu'un intrus peut déduire un attribut avec un niveau de confiance élevé. En général, ce type de divulgation n'est pas pris en compte dans la protection d'un jeu de données.

Pour respecter *stricto sensu* la réglementation, une approche consiste à distinguer différents types d'utilisateurs. Les utilisateurs standard auront accès à moins d'informations (moins de

1. En dehors de l'Europe des textes équivalents existent, comme l'*Australian Privacy Act* de 1988.
2. CLIFTON et al. 2012 proposent une classification des différents risques de divulgation.

variables ou modalités plus larges), tandis que des utilisateurs spécifiques (chercheurs), bénéficieront d'un accès restreint à davantage de données, accessibles au moyen de serveurs sécurisés, à condition qu'ils justifient au préalable leur demande et respectent des procédures.

Une approche complémentaire consiste à introduire de la perturbation dans les données diffusées, afin de ramener le risque de divulgation à un niveau acceptable. On met alors en place une méthode SDC, ce qui revient à réduire l'utilité des données en échange d'en augmenter la protection. Traditionnellement, les méthodes SDC ne tiennent pas compte des caractéristiques spatiales. Il se peut donc que les corrélations spatiales soient largement déformées avant et après la perturbation. La sous-section suivante recense des arguments en faveur de stratégies de confidentialité prenant en compte la géographie.

14.1.2 Spécificité des données spatiales

Les experts ayant à traiter la confidentialité de données spatiales sont confrontés à un paradoxe. D'un côté, ces données ont besoin de davantage de protection parce qu'elles pourraient permettre plus d'identifications, mais de l'autre, elles ouvrent de nombreuses possibilités d'analyse, que les utilisateurs souhaitent conserver.

Considérations théoriques

Dans le manuel d'Eurostat qui rassemble des consignes en matière de confidentialité (HUNDEPOOL et al. 2010), trois niveaux différents de quasi-identifiants sont suggérés. Seul le lieu géographique est pris en compte dans la catégorie des variables dites "extrêmement identifiantes". En effet, le risque de divulgation est plus élevé lorsque l'on considère des données spatiales, pour plusieurs raisons.

Premièrement, le risque de divulgation d'identité augmente en présence de données spatiales, car il est plus facile de mobiliser des connaissances personnelles. En effet, parmi les caractéristiques potentiellement partagées avec un individu (âge, genre, etc.), l'appartenance à un même voisinage est sans doute celle qui augmente le plus la probabilité de le connaître personnellement. En outre, l'identification des adresses est devenue possible avec le développement du *web scraping* ou d'outils en accès libre tels que *Google Earth*, qui rendent possibles la rétro-ingénierie (CURTIS et al. 2006) ou l'identification directe (ELLIOT et al. 2014). La densité de population est donc un prédicteur fondamental du risque de divulgation : plus la densité est basse, plus le risque de divulgation est élevé.

Deuxièmement, le risque de divulgation d'attributs augmente en présence de données spatiales, en raison de la première loi de la géographie de Tobler, selon laquelle : "*tout interagit avec tout, mais deux objets proches ont plus de chances d'interagir que deux objets éloignés*". Par conséquent, le degré de dissimilarité d'un individu par rapport à ses voisins est un autre bon prédicteur du risque de divulgation.

Enfin, le risque de divulgation est aussi plus élevé pour les données spatiales à cause des questions de différenciation. Lorsque des données sont diffusées dans différentes géographies (typiquement, des frontières administratives d'une part et des grilles de carreaux d'autre part), dans certains cas, on peut déduire les caractéristiques d'un individu par soustraction. Toute personne maîtrisant les systèmes d'information géographique devient donc un intrus potentiel. Cette question spécifique de la différenciation géographique fera l'objet de la section 14.4.

Considérations techniques

Techniquement, le maillage de diffusion (zonages, contours administratifs ou grilles de carreaux) est une variable catégorielle comme une autre (une dimension supplémentaire pour les données tabulées). Il est donc possible, avec un logiciel classique, de traiter le risque de divulgation sans aucune considération géographique, simplement en considérant le maillage comme une variable

présentant de nombreuses modalités. Un traitement prenant en compte la géographie préserverait les phénomènes spatiaux sous-jacents, mais aucun logiciel spécifique n'a été développé pour l'instant.

Sur un plan purement pratique, le traitement de données spatiales ajoute une couche de complexité dans le processus de contrôle de la confidentialité, parce qu'il nécessite des puissances de calcul importantes. En particulier, certaines méthodes prétabulées impliquent de spécifier la structure du voisinage avec une matrice de poids (appelée "matrice W"), dont la dimension peut facilement devenir ingérable pour les ordinateurs classiques. Sur les données tabulées également, la détection des problèmes de différenciation requiert parfois de croiser de nombreuses dimensions (problème NP-difficile).

Une préoccupation grandissante

Enfin, et surtout, les données spatiales deviennent de plus en plus nombreuses et populaires, en particulier sous la forme de données carroyées. La diffusion croissante de données carroyées (aux niveaux national ou international [3]) est rendue possible par une géolocalisation de plus en plus systématique des données par les INS.

Les données carroyées ont de nombreux avantages. Elles répondent bien au besoin de meilleure représentation des réalités socio-économiques en s'affranchissant de tout zonage administratif, qui ne rend compte ni des réalités socio-économiques, ni des réalités naturelles (CLARKE 1995, DEICHMANN et al. 2001). Elles décrivent également mieux les régions faiblement peuplées, comme en Finlande ou en Suède (TAMMILEHTO-LUODE 2011). Comme les carreaux ont toujours la même taille, les données carroyées garantissent une comparabilité entre les territoires et dans le temps. Si nécessaire, les carreaux peuvent être agrégés pour former des zones d'étude à façon. Les données carroyées constituent aussi une bonne source d'information auxiliaire notamment à des fins d'échantillonnage local. Enfin, il est facile de leur intégrer d'autres données de différentes natures, avec des utilisations possibles dans de nombreuses disciplines : météorologie, environnement, santé, télécommunications, marketing, etc.

La section suivante présente comment, dans ce contexte, il est possible d'introduire des aspects géographiques dans les traitements de la confidentialité dans le but de conserver une utilité maximale des données.

14.1.3 Recommandations pour évaluer le risque de divulgation

L'évaluation quantitative du risque de divulgation est une étape cruciale pour les experts en confidentialité. Des indicateurs du risque de divulgation ont été proposés dans le contexte des données non spatiales (WILLENBORG et al. 2012, DUNCAN et al. 2001, DOYLE et al. 2001). Ils sont souvent fondés sur la théorie de la décision (LAMBERT 1993, DUNCAN et al. 2001). Pour décrire un jeu de microdonnées, le k-anonymat et la l-diversité sont des critères couramment utilisés. Un jeu de données satisfait le k-anonymat si, pour chaque combinaison de modalités de quasi-identifiants, il y a au moins k observations. Il satisfait la l-diversité lorsque, pour chaque combinaison de modalités de quasi-identifiants, il y a au moins l modalités "bien représentées" pour les variables sensibles. La l-diversité étend le k-anonymat en assurant une hétérogénéité intra-groupe des variables sensibles, afin d'éviter une divulgation d'attributs par homogénéité trop grande d'un groupe.

En présence de données spatiales, on peut calculer des scores individuels de risque pour prendre en compte le fait qu'une observation est exposée au risque de divulgation. La tâche n'est toutefois pas facile, et il n'existe pas de mesure binaire consensuelle du fait d'être à risque ou non.

3. Dans les années 1990, le projet *Gridded Population of the World* a commencé à parler de données carroyées à l'échelle mondiale. Il a été suivi par une amélioration continue de la résolution de la grille (DEICHMANN et al. 2001). Début 2010, le projet Geostat a été lancé (coopération entre Eurostat et le Forum Européen sur la Géographie et la Statistique (EFGS)). La première partie de Geostat concernait spécifiquement les données carroyées (BACKER et al. 2011), tandis que la seconde partie visait à encourager l'intégration de l'information géographique, dans l'objectif de mieux décrire et analyser la société et l'environnement (HALDORSON et al. 2017).

Que les données soient spatiales ou non, une approche post-tabulée consiste à construire les données tabulées comme elles seraient diffusées sans traitement de confidentialité, et de marquer les cellules qui ne respectent pas les contraintes (effectif sous un seuil, règle de prédominance - également appelée règle (n,k), règle des p %[4]). Les observations à risque sont alors toutes celles qui se trouvent dans ces cellules à risque. Dans le cas des données spatiales, les observations à risque peuvent être exhibées selon les mêmes règles, en considérant que la maille géographique est une variable comme une autre des données tabulées.

Une autre approche (prétabulée) consiste à travailler directement à partir des microdonnées. On associe à chaque observation la probabilité d'être réidentifiée par un intrus, avec l'idée que le risque d'une observation est élevé s'il n'y a pas d'observations similaires dans son voisinage. On calcule donc un score pour chaque observation, qui indique la probabilité de trouver dans le voisinage une autre observation partageant les mêmes modalités pour un ensemble de quasi-identifiants. Un individu vivant seul dans une zone vide sera bien considéré comme exposé au risque de divulgation, mais un individu âgé vivant au milieu d'une population majoritairement jeune le sera également.

Dans l'idéal, un tel score impliquerait de définir un voisinage entre deux observations (distance euclidienne, nombre de ménages dans un disque, voisinage de Moore ou de von Neumann[5]), et de constituer une matrice $n \times n$ à partir des données exhaustives[6]. Cependant, pour les régions peuplées, les puissances de calcul sont actuellement limitantes pour de tels calculs. Pour pallier ces problèmes, on peut fonder l'évaluation du risque sur les éléments suivants :

— comptages des fréquences des variables sensibles (voir également algorithme des *special unique* développé dans ELLIOT et al. 2005) ;
— définition simplifiée du voisinage en considérant l'appartenance à une même zone de niveau hiérarchique supérieur. Cela suppose de disposer d'un système de géographies imbriquées[7]. On n'utilise alors pas directement l'information géographique.

Deux exemples permettant de cibler les observations les plus à risque sont présentés ci-après.

Encadré 14.1.1 Dans SHLOMO et al. 2010, un score est calculé pour chaque observation, comme suit. M variables-clé (ou quasi-identifiants, tous catégoriels) sont choisies, chacune ayant k_m modalités ($m = 1, ..., M$). On considère un système hiérarchique de niveaux géographiques (par exemple, la partition en NUTS imbriqués, ou un carroyage composé de carreaux de différentes tailles). Pour chaque niveau géographique l avec G modalités ($g = 1, ..., G$, par exemple G carreaux), on définit le comptage univarié $N_k^{g,m}$ ($k = 1, ..., k_m$). Le tableau $N_k^{g,m}$ ci-dessous a donc $G * \sum_{m=1}^{M} k_m$ cellules.

g	Mod. A1	Mod. A2	Mod. A3	Mod. B1	Mod. B2
1	5	4	1	7	3
2	4	3	3	9	1
...					
G	5	0	5	6	4

Pour chaque niveau géographique l (par exemple le carreau), on calcule pour chaque observation i (portant les modalités $k_1^i, ..., k_M^i$ et appartenant à la maille g^i) un score égal à la moyenne

4. Toutes ces règles sont bien connues dans la littérature générale sur le contrôle de la confidentialité et ne sont donc pas développées ici.
5. Se référer à la lecture du chapitre 2 : "Codifier la structure de voisinage".
6. Ici n est le nombre d'observations dans les microdonnées.
7. Ce système hiérarchique peut être la grille de diffusion finale, ou être spécifié de façon *ad-hoc*.

des inverses des fréquences :

$$R_i^l = \frac{\sum_{m=1}^{M} 1/N_{k_m^i}^{g_i,m}}{M}. \tag{14.1}$$

Dans l'exemple ci-dessus, un individu i de la région g=1 avec les modalités ($A1$, $B1$) a un score égal à $(1/5 + 1/7)/2 \simeq 0.17$. Des seuils T^l sont ensuite fixés pour chaque niveau géographique l, et des scores supérieurs à ces seuils indiquent les observations exposées au risque de divulgation. Les seuils correspondent en général à des quantiles ; ils sont définis par l'expert qui décide quelle proportion de la population il convient de considérer à risque. Le problème est que ces seuils sont *data-specific* : un seuil peut être pertinent pour un certain maillage mais pas pour un autre. Par exemple, 10 % de carreaux à risque ne signifie pas la même chose si le carreau fait 10 mètres ou 10 kilomètres de côté.

Des données carroyées du recensement hongrois ont été confidentialisées avec la même approche pour cibler les individus à risque, mais en introduisant cette fois des distributions multivariées. Ainsi dans NAGY 2015, des *flag values* sont calculées pour chaque combinaison possible de 3 attributs (dont le carreau), et les n observations les plus à risque seront les n premières, en triant par ordre décroissant de la somme de ces *flag values*. Le nombre de cellules du tableau de fréquences est alors de $G * \prod_{m=1}^{M} k_m$ cellules (tableau creux). Si M est élevé (ou si la plupart des k_m sont élevés), des limites de coût computationnel sont possibles. Une solution peut consister à créer des quasi-identifiants *ad hoc* qui croisent des variables bien choisies, ou d'ajouter *a posteriori* à l'échantillon à risque les observations aux combinaisons de modalités très rares (par exemple veuves âgées de moins de 20 ans, etc.).

Dans ces deux exemples cités, les ménages à risque sont définis comme ceux comprenant au moins un individu à risque.

14.2 Comment gérer le risque de divulgation ?

Une fois les observations à risque identifiées, on souhaite leur faire subir une perturbation, afin de ramener le risque global du fichier à un niveau acceptable. La section 14.2.1 énonce des éléments généraux sur les méthodes SDC, tandis que la section 14.2.2 présente celles qui prennent en compte la spécificité des données spatiales. Pour finir, la section 14.2.3 propose des indicateurs permettant d'évaluer l'efficacité d'une méthode.

14.2.1 Méthodes prétabulées et post-tabulées

Traditionnellement, dans la littérature traitant du contrôle de la confidentialité et des méthodes dites SDC, on distingue les méthodes post-tabulées, appliquées aux tableaux (hypercubes), et les méthodes prétabulées, appliquées aux microdonnées. Dans le cas du recensement, en pratique, la plupart des pays adopte des méthodes post-tabulées, par exemple en regroupant des cellules jusqu'à atteindre des seuils suffisants. Ces méthodes doivent être appliquées plusieurs fois (autant que de tableaux différents à diffuser), ce qui peut devenir très lourd lorsque différentes géographies sont utilisées ou que l'on souhaite conserver une cohérence entre différents tableaux liés. En outre, les méthodes post-tabulées peuvent fausser les corrélations entre variables (KAMLET et al. 1985) et les corrélations spatiales.

Les méthodes prétabulées apparaissent alors comme une solution intéressante [8]. Premier avantage, ces méthodes ne sont appliquées qu'une seule fois : en effet, si les microdonnées sont protégées, toutes les agrégations possibles à partir de ces microdonnées le seront également, et la cohérence

8. L'objectif de ces méthodes n'est pas de publier les microdonnées elles-mêmes, mais de produire un fichier de microdonnées qui sera commun aux données tabulées ou carroyées.

entre les différents tableaux est préservée. Ensuite, elles sont très largement paramétrables [9] et permettent ainsi une grande flexibilité des produits statistiques, qu'il s'agisse de données carroyées ou d'hypercubes (voire des données à façon pour des besoins personnalisés d'utilisateurs). Un autre avantage est que certaines méthodes prétabulées (comme le *swapping*) peuvent être non biaisées, là où la plupart des méthodes post-tabulées impliquent de supprimer des cellules et donc d'introduire un biais dans l'estimation de paramètres, voire de rendre impossible leur estimation. Malgré ces avantages, en pratique, il n'est pas réaliste d'envisager un fichier unique à partir duquel il serait possible d'extraire tous les tableaux à diffuser en toute sécurité, car cela impliquerait de toucher à trop d'observations (YOUNG et al. 2009). En outre, les méthodes prétabulées ont l'inconvénient de laisser croire aux utilisateurs que rien n'est fait pour garantir la confidentialité (LONGHURST et al. 2007, SHLOMO 2007). En effet, avec uniquement des méthodes prétabulées, on continuerait de diffuser des cellules à effectifs très faibles.

Un compromis classique consiste alors à mettre en œuvre un premier niveau de protection dans le fichier de microdonnées, puis un deuxième niveau de protection dans les tableaux (MASSELL et al. 2006, HETTIARACHCHI 2013). Cela permet de garantir le respect de règles empiriques souvent souhaitées (seuils de diffusion, règle (n,k), règle p %, etc.). Par exemple, après perturbation des microdonnées, les cellules qui ne contiennent qu'un ménage pour une variable donnée seront supprimées.

Les méthodes prétabulées semblent plus appropriées à la prise en compte de l'information géographique, dans le sens où l'on peut l'utiliser directement pour cibler les observations les plus à risque, en vue de leur faire subir une perturbation.

14.2.2 Méthodes de protection prenant en compte la géographie

Les méthodes SDC traditionnelles font déjà l'objet d'un manuel dédié d'Eurostat (HUNDEPOOL et al. 2010, HUNDEPOOL et al. 2012) et ne sont donc pas détaillées dans le présent chapitre. Nous évoquons ici des méthodes qui prennent plus explicitement en compte l'information géographique.

Imputation locale

MARKKULA 1999 est l'un des premiers articles qui tienne compte de la géographie dans le choix de méthode. Sa méthode, l'imputation restreinte locale (*Local Restricted Imputation*, LRI), a été co-développée par l'INS de Finlande et l'Université de Jyvaskyla, et a été testée sur les données du recensement finlandais. Elle comprend trois phases :

1. définition du cadre : seuil de confidentialité et configuration spatiale (en l'occurrence 3 niveaux de géographies imbriqués) ;
2. identification des zones à risque, c'est-à-dire dont le nombre d'individus est en dessous du seuil ;
3. imputation des nouvelles valeurs pour les zones à risque. Deux techniques sont étudiées : imputation par la moyenne de toutes les zones à risque de la même zone du niveau hiérarchique supérieur, ou imputation par un système de permutation aléatoire avec la valeur d'une autre zone à risque sélectionnée aléatoirement dans le niveau hiérarchique supérieur.

La méthode LRI vise principalement à préserver les relations spatiales, tout en restituant autant d'utilité que possible. Elle peut également être adaptée aux données carroyées (TAMMILEHTO-LUODE 2011). Elle a l'avantage d'être simple à comprendre et d'être consistante : les totaux du niveau hiérarchique supérieur sont conservés. Cependant, la documentation sur la méthode est insuffisante pour la reproduire précisément.

9. En général, les INS ne révèlent pas aux utilisateurs le jeu de paramètres choisi (taux d'observations échangées, matrices de transition dites PRAM, paramètre de lois de distribution, etc.), afin de limiter le risque de rétro-ingénierie (SHLOMO et al. 2010, ZIMMERMAN et al. 2008).

Agrégation géographique

Le plus souvent, la règle de protection des données n'est autre qu'un seuil, en deçà duquel on interdit la diffusion de statistiques. Dans le cas des données carroyées, une stratégie peut consister à assembler des carreaux contigus en plus gros polygones (par exemple des rectangles ou des carrés plus gros), de sorte que chaque polygone respecte le seuil. Dans ces méthodes, on cherche la grille de diffusion optimale, permettant de diffuser des données au niveau le plus fin possible. On se situe donc à la frontière entre les méthodes SDC et la visualisation des données, puisqu'en pratique, on crée des cartes de résolutions diverses. Les nouveaux polygones peuvent être obtenus par agrégation (regroupement de carreaux jusqu'à atteindre le seuil) ou par désagrégation (on part d'un grand carreau et on le découpe jusqu'à ce qu'il ne soit plus possible de le découper sans passer sous le seuil). Ces méthodes présentent des propriétés intéressantes : l'additivité est préservée, ainsi que les moyennes par polygone. De plus, cette méthode n'introduit pas de "faux zéros" et respecte, par sa construction, la règle du seuil.

En revanche, l'agrégation géographique ne résout pas le problème de réidentification aisée des valeurs extrêmes ou des combinaisons rares. Par ailleurs, les polygones ne correspondent pas toujours à une réalité géographique : par exemple, une île peut être associée à la terre la plus proche. Enfin, elle donne également lieu à des problèmes de différenciation avec les autres niveaux de diffusion (coexistence de frontières géométriques et administratives).

Deux versions de la méthode d'agrégation géographique sont présentées ci-dessous : la première mise au point par l'Insee pour la diffusion de statistiques issues de la source fiscale en 2013, et la deuxième pour une représentation des données du parc immobilier en Allemagne (BEHNISCH et al. 2013).

Encadré 14.2.1 — Exemple 1 : carroyage en rectangles. En 2013, pour publier des statistiques de revenus dans une grille de carreaux de 200 mètres de côté, l'Insee a dû se plier au seuil réglementaire : aucune statistique issue de la source fiscale ne peut être publiée si elle ne se rapporte pas à un minimum de 11 ménages fiscaux. Pour ce faire, l'Insee a choisi d'utiliser un algorithme de désagrégation. Le territoire métropolitain est au préalable divisé en 36 gros carreaux de dimensions similaires. Chacun de ces carreaux est ensuite découpé horizontalement ou verticalement, formant deux sous-rectangles. Ces derniers sont découpés de nouveau horizontalement ou verticalement, et ainsi de suite. Les découpages horizontaux ou verticaux passent toujours par le centre de gravité du rectangle, pondéré par la population.

À chaque étape, on opte pour le découpage horizontal, le découpage vertical ou l'absence de découpage, comme suit (voir figure 14.1) :

— si les deux découpages (horizontal et vertical) produisent au moins un des sous-rectangles comportant moins de 11 ménages fiscaux, le découpage n'a pas lieu ;
— si l'un des deux découpages seulement produit deux sous-rectangles d'au moins 11 ménages fiscaux chacun, c'est ce découpage qui est choisi ;
— si les deux découpages produisent chacun deux sous-rectangles de plus de 11 ménages fiscaux, on choisit celui qui minimise la somme des dispersions des deux sous-rectangles. La dispersion d'un rectangle est évaluée par la somme des carrés des distances entre son centre de gravité et ceux de ses carreaux habités, pondérés par la population des carreaux.

La grille de rectangles a l'avantage de bien rendre compte de l'irrégularité des données, mais a l'inconvénient d'être conçue pour un jeu de données particulier : la grille ne convient ni à une autre source, ni à un autre millésime de la même source.

Encadré 14.2.2 — Exemple 2 : méthode du *quadtree*. La méthode dite *quadtree* est un autre algorithme d'agrégation géographique qui permet de représenter des données de multiples

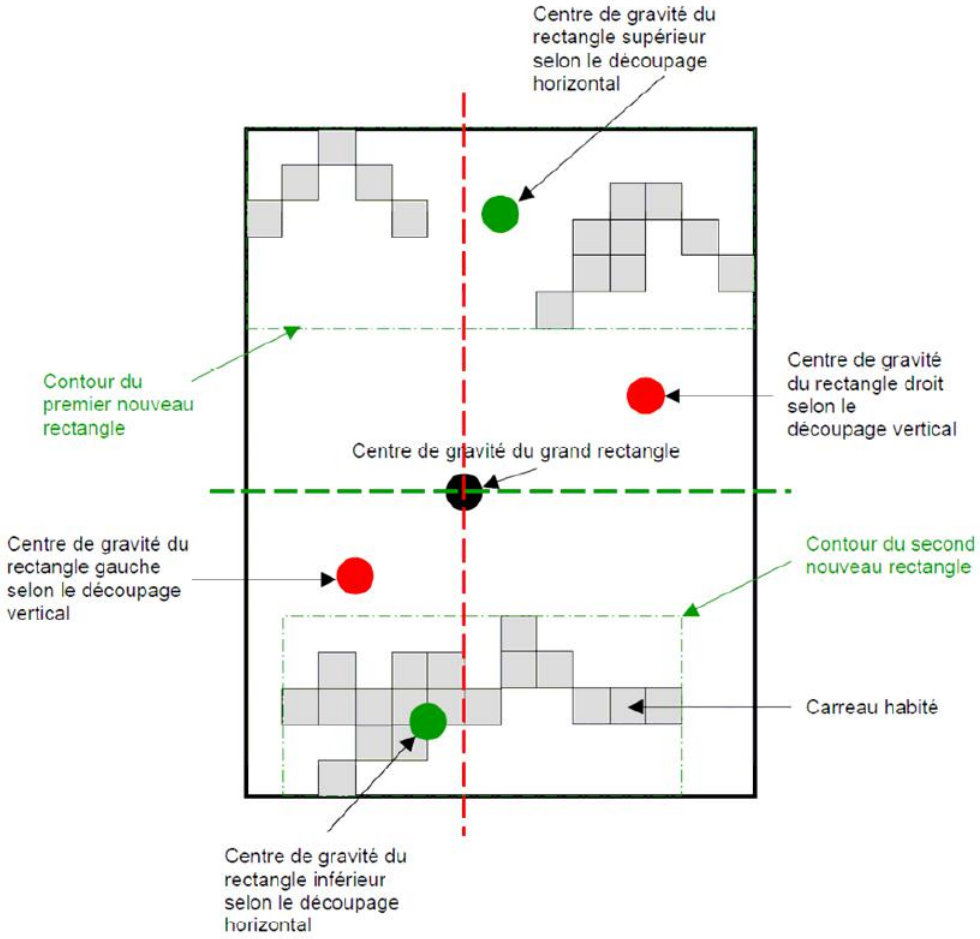


FIGURE 14.1 – Exemple de compromis entre découpage horizontal et vertical

résolutions en une seule visualisation. Elle a été mise en œuvre par l'Institut de Leibniz pour le développement régional et urbain écologique (BEHNISCH et al. 2013) pour un projet de représentation du parc immobilier en Allemagne. L'algorithme est initialisé avec la résolution de grille la plus fine (par ex. 250 m × 250 m) et, si un carreau contient moins d'unités que le seuil, il est regroupé avec ses voisins pour former un carreau plus gros (500 m × 500 m). L'algorithme s'arrête lorsque tous les carreaux sont au-dessus du seuil (voir figure 14.2).

La méthode *quadtree* permet d'obtenir des grilles consistantes d'une source à l'autre, mais également entre différents millésimes d'une même source. Autrement dit, il est possible de trouver un maillage pour lequel différentes sources peuvent être croisées, à des fins d'analyse. Toutefois, cette méthode a l'inconvénient de masquer certains carreaux supérieurs au seuil (en gras sur la figure 14.2), et de ne pas totalement résoudre le MAUP (*Modifiable Areal Unit Problem*, problème d'agrégation spatiale), puisque les données sont diffusées dans une grille définie de manière déterministe.

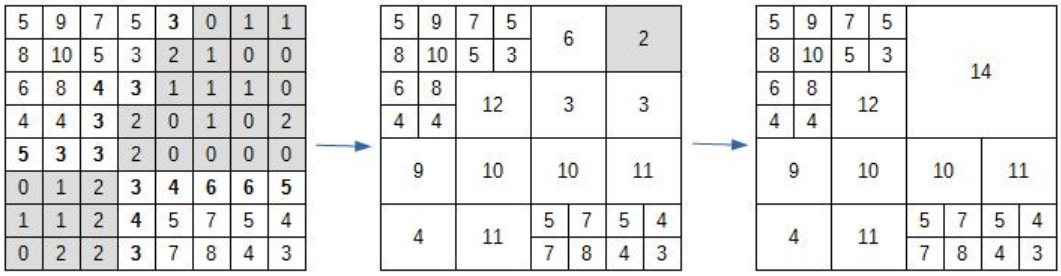


FIGURE 14.2 – Exemple de méthode *quadtree* (agrégative) appliquée à des données carroyées, avec un seuil de 3

Targeted record swapping

De façon générale, le *swapping* (parfois considéré comme un cas particulier de méthodes PRAM, pour *Post RAndomisation Method*, GOUWELEEUW et al. 1998) consiste à permuter les attributs de deux observations entre elles. La méthode de *Targeted Record Swapping* (TRS), par opposition au *random swapping*, vise à permuter des attributs des observations les plus exposées au risque. Cette méthode prétabulée est souvent présentée comme offrant un bon compromis entre risque et utilité. Le *swapping* produit des données consistantes, puisqu'une observation prend la place d'une autre ; quelles que soient les variables considérées, les distributions univariées sont conservées, et le nombre d'observations par cellule ou carreau n'est pas modifié (plus particulièrement, l'échange n'introduit pas de "faux zéros" dans les fréquences). Les inconsistances n'apparaissent que lorsque l'on croise plusieurs dimensions.

L'institut de statistique britannique (ONS) est à l'origine d'une littérature autour des méthodes prétabulées de type *swapping* et d'extensions prenant en compte la géographie, appliquées à la diffusion de données de recensement (BROWN 2003, SHLOMO 2007, YOUNG et al. 2009, SHLOMO et al. 2010). Ainsi, le TRS a été testé sur des données synthétiques issues du recensement britannique. Au Japon, ITO et al. 2014 ont également testé un algorithme de TRS pour la diffusion de données du recensement de 2005. Le principal apport du TRS est de cibler pour la perturbation les observations qui présentent un risque de réidentification élevé. Ce risque dépend de la grille de diffusion. La méthode fait en sorte que les observations permutées ne soient pas trop distantes géographiquement.

Les premières versions du TRS ont été mises au point pour des géographies imbriquées (BROWN 2003, SHLOMO 2005), ou pour des données carroyées (NAGY 2015). Dans ces différents travaux, deux individus ne peuvent pas être échangés s'ils n'appartiennent pas à la même zone de niveau

hiérarchique supérieur. Pour une observation à risque donnée, les observations éligibles sont celles qui appartiennent au même voisinage, et parmi elles, on sélectionne celle la plus proche au sens d'un ensemble de variables-clé, en priorisant les autres observations à risque et en éliminant celles qui ont déjà fait l'objet d'une permutation.

La méthode de *Local Density Swapping* (LDS), décrite dans YOUNG et al. 2009, va plus loin en utilisant directement les coordonnées géographiques pour définir le voisinage. Dans cette méthode, les observations éligibles pour la permutation sont celles qui ont les mêmes valeurs pour un certain nombre de variables-clé d'appariement et, parmi elles, on choisit l'observation qui minimise une fonction de distance. Le cœur de la méthode LDS est de remplacer la distance euclidienne par le nombre d'observations situées "entre" les deux observations à permuter, c'est-à-dire se trouvant à l'intérieur du disque dont le centre se trouve sur l'observation d'origine, et le rayon est le segment entre les deux observations. Cela permet de tenir compte de la densité de population. Comme précédemment, la priorité est donnée aux observations qui n'ont pas déjà été échangées.

La méthode LDS est très flexible puisqu'elle est largement paramétrable (nombre d'observations à permuter, choix de la distance, liste des variables-clé d'appariement). Elle est également particulièrement adaptée au contexte des données carroyées, puisqu'elle prend plus finement en compte la géographie que les autres méthodes présentées. Toutefois, comme toute méthode prétabulée, elle a l'inconvénient de ne pas être suffisante, et de laisser croire au public que rien n'a été fait pour garantir la confidentialité.

Extensions

Trajectoires

Les données de trajectoire peuvent être considérées comme un cas particulier de données spatiales. De nombreuses technologies permettent de collecter des données bilocalisées. Or, celles-ci sont très sensibles, car elles sont éloquentes quant aux habitudes individuelles (lieux souvent fréquentés). C'est pourquoi leur dé-identification est plus délicate. Une trajectoire est souvent associée à une dimension temporelle, qu'il est pertinent de prendre en compte dans la méthode de protection. Ces deux aspects (temporel et spatial) peuvent être mobilisés pour définir la distance entre deux trajectoires.

DOMINGO-FERRER et al. 2011 présentent deux méthodes pour anonymiser les données de trajectoires, appelées *SwapLocations* et *ReachLocations*. La première conserve le k-anonymat de la trajectoire, tandis que la seconde garantit la l-diversité des positions géographiques.

SHLOMO et al. 2013 suggèrent un protocole de détection et de correction des données aberrantes de trajectoires, qui tient compte des informations géographiques. Les auteurs s'intéressent aux trajets domicile-travail, caractérisés par deux positions géographiques et un temps de trajet (en minutes). Les trajectoires aberrantes sont définies par mode de transport. Dans une première étape, les trajectoires aberrantes sont détectées, en utilisant la distance de Mahalanobis (hypothèse de distribution normale multivariée). Dans une seconde étape, ces trajectoires aberrantes sont modifiées : on déplace le lieu du domicile en laissant le lieu de travail inchangé, afin de ne pas introduire d'incohérence (la perturbation serait facile à détecter si une usine était placée là où il n'en existe pas en réalité). Pour ce faire, les auteurs définissent une fonction de cohérence, qui évalue si une trajectoire est plausible au regard de ce qui est observé dans les trajectoires non aberrantes. L'article teste plusieurs algorithmes :

— *swapping* : algorithme itératif qui, pour chaque sous-groupe mode de transport $\times$ genre $\times$ âge, crée des paires de trajectoires et permute les lieux de résidence au sein des paires, sans toucher aux lieux de travail. À chaque itération, le principe est d'apparier les observations qui optimisent la fonction de cohérence. On arrête les itérations lorsque le gain de cohérence devient négligeable ;
— *hot deck* : plutôt que d'être échangés, les lieux de résidence des trajectoires aberrantes sont effacés et remplacés par imputation par la valeur d'un "donneur" qui présente les mêmes

caractéristiques. La sélection du donneur peut être effectuée en maximisant la cohérence parmi tous les donneurs potentiels d'un voisinage, ou en minimisant la différence en termes de temps de trajet.

Le *hot deck* corrige davantage d'observations aberrantes que le *swapping*, mais le *swapping* limite la perte d'informations. Dans les deux cas, il se peut que des observations non aberrantes le deviennent (mais cela est moins fréquent avec le *swapping*).

Geomasking

Le terme de *geomasking* a été introduit par ARMSTRONG et al. 1999. Il désigne l'ensemble des méthodes modifiant les positions géographiques dans des données ponctuelles, afin de leur apporter davantage de protection. L'une des techniques de *geomasking* est la méthode dite du "donut", dans laquelle chaque adresse géolocalisée est relocalisée dans une direction aléatoire, à une distance comprise entre un minimum et un maximum.

Le *geomasking* a peu fait l'objet d'applications en économie, mais est largement utilisé en épidémiologie ou pour des données de criminalité. Dans ces domaines, il s'agit de diffuser des données ponctuelles, alors que dans ce chapitre, l'enjeu est de diffuser des données dans un maillage (grille régulière de carreaux ou zonage administratif) : ici, les microdonnées ne sont pas le produit définitif, mais un produit intermédiaire que l'on altère pour atteindre le produit final. Dans le cas du recensement, on exclut en général de relocaliser les ménages, car il pourrait résulter de cela des incohérences évidentes (par exemple, un ménage risquerait de se retrouver au beau milieu d'un lac). Un autre inconvénient serait aussi de créer des "faux zéros" et de ne pas conserver les "vrais zéros".

14.2.3 Comment évaluer l'efficacité d'une méthode ?

Indicateurs spatiaux d'utilité

L'application d'une méthode SDC consiste à détériorer l'utilité des données en échange d'une meilleure protection, ce qui se traduit par une perte d'information pour les utilisateurs. Pour arbitrer entre les différentes méthodes SDC, mesurer l'utilité revient en fait à mesurer une désutilité ou une distorsion. D'après WILLENBORG et al. 2012 au sujet de l'incidence des méthodes SDC sur les microdonnées, les pertes d'information sont de deux types : augmentation de la variance dans l'estimation d'un paramètre d'une part, et introduction d'un biais d'autre part (ce qui est évidemment le cas lorsque l'on masque les valeurs extrêmes).

Différents indicateurs mesurent la perte d'information (DOMINGO-FERRER et al. 2001), au premier rang desquels la part d'observations perturbées. Mais on peut citer également :

— pour les variables continues : l'erreur quadratique moyenne, l'erreur absolue moyenne, les changements de rang moyens (une fois les observations triées selon la variable en question) ou la comparaison du coefficient de Pearson entre deux variables dont on sait qu'elles sont corrélées. Si la source à diffuser vise principalement à produire un indicateur particulier tel que le taux de chômage, il est également pertinent de vérifier si un biais n'a pas été introduit entre le fichier d'origine et le fichier anonymisé. D'autres métriques fondées sur des modèles peuvent enfin être utilisées, par exemple en calculant si deux intervalles de confiance se chevauchent pour une même régression logistique, dans les deux fichiers (original et modifié) (DE WOLF 2015) ;
— pour les variables catégorielles : comparaison directe des fréquences, mesures fondées sur l'entropie comme la distance de Hellinger (TORRA et al. 2013), ou comparaison de tableaux de contingence entre deux variables que l'on sait corrélées.

Pour les données spatiales, on peut ajouter à cette liste la proportion d'unités géographiques ayant subi une perturbation, l'écart moyen absolu (AAD) des comptages d'un attribut donné, avant et après perturbation, calculé au niveau de la maille (ou du carreau), ou encore les indicateurs

locaux d'association spatiale (LISA) ou indices de Moran [10], pour une variable dont on sait qu'elle présente une dépendance spatiale.

Cartes risque-utilité

Afin de comparer différentes stratégies de protection (choix d'une méthode SDC ou choix du jeu de paramètres adéquat pour une méthode SDC choisie), il est recommandé de tracer des cartes risque-utilité (ou *R-U maps*) pour différents niveaux de risque.

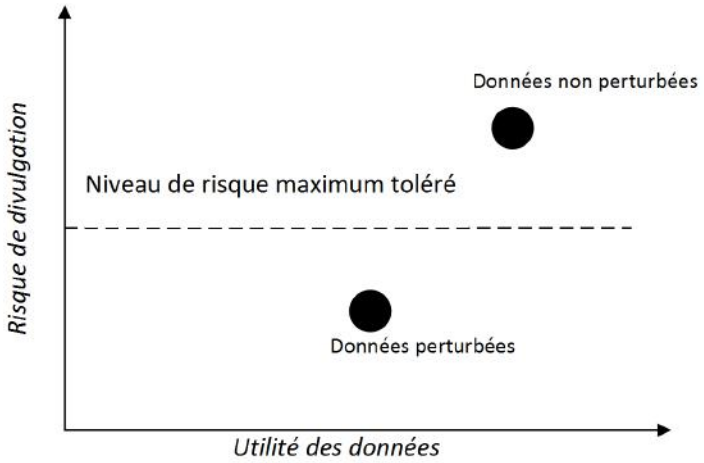


FIGURE 14.3 – Principe général d'une carte risque-utilité

Les *R-U maps* (figure 14.3) ont été formalisées pour la première fois par DUNCAN et al. 2001 (pour une méthode d'ajout de bruit), et ont ensuite été utilisées dans de nombreux articles (YOUNG et al. 2009, CLIFTON et al. 2012, GOMATAM et al. 2005). Elles constituent un outil pratique formalisant la prise de décisions et proposant une représentation synthétique du compromis à trouver entre le risque de divulgation noté R, que l'on souhaite faible, et l'utilité des données notée U, que l'on souhaite élevée. Une *R-U map* est un schéma qui représente comment évoluent R et U lorsque l'on modifie la méthode SDC ou les paramètres d'une même méthode.

14.3 Application à une grille de carreaux de 1 km²

En 2017, le programme d'Eurostat "Protection harmonisée des données de recensement du système statistique européen [11]" avait pour objectif d'harmoniser les stratégies de confidentialité des données des différents recensements européens, qu'il s'agisse des hypercubes ou des données carroyées. Dans ce cadre, deux méthodes SDC, jugées complémentaires, ont été mises en avant, car présentant un bon compromis entre risque et utilité. Dans une première étape, on produit un fichier de microdonnées perturbé, par exemple en mettant en œuvre la méthode du TRS. Dans une deuxième étape, on produit les données carroyées et les données tabulées à partir de ce fichier perturbé, et on applique un niveau supplémentaire de protection, par exemple en ajoutant de bruit sur les cellules comme le propose la *cell-key method*, qui s'inspire de travaux de l'*Australian Bureau of Statistics* (FRASER et al. 2005).

Dans la section suivante, nous nous focalisons sur le TRS proposé en première étape, et nous tentons de déterminer dans quelle mesure cette méthode dégrade, ou au contraire préserve, les corrélations spatiales. Nous nous appuyons sur les données fiscales d'une région française de petite

10. Voir chapitre 3 : "Indices d'autocorrélation spatiale".
11. SSE

taille. Nous présentons les principales étapes de cette méthode et les résultats grâce à une analyse risque-utilité.

14.3.1 *Targeted record swapping* : détails de la méthode

Les choix d'implémentation de la présente application sont largement inspirés d'un programme de l'ONS [12], qui a été adapté afin de mieux s'ajuster aux données françaises. L'algorithme d'origine s'applique à des données "hiérarchiques", structurées en 3 niveaux imbriqués ($niveau1 \subseteq niveau2 \subseteq niveau3$). La méthode comprend quatre étapes, détaillées ci-dessous.

Étape 1 : Ciblage des observations à risque

La première étape consiste à identifier les observations ayant le plus besoin d'être perturbées. Un individu est considéré ou non comme exposé au risque au sens d'un ensemble de caractéristiques donné. Être exposé au risque signifie que les autres observations similaires sont très rares dans le voisinage : un score de rareté est calculé pour chaque individu, comme suggéré plus haut (moyenne des inverses des fréquences) et les individus dont le score est supérieur à un seuil (quantile) sont marqués comme à risque. Les ménages à risque sont ensuite définis comme ceux qui comportent au moins un individu à risque.

On associe à chaque individu un niveau géographique de risque : si la modalité est très rare même au niveau hiérarchique supérieur (moins de X individus partageant la même modalité dans la zone), l'individu est "unique" pour ce niveau géographique. Le niveau géographique de risque du ménage est défini comme étant le niveau le plus élevé parmi tous les individus qui le composent. Un ménage que l'on considère à risque au niveau 2 pourra être apparié avec un ménage plus éloigné que s'il avait été à risque seulement au niveau 1.

Étape 2 : Constitution d'un échantillon

Le principe de cette étape est de constituer un échantillon de ménages, dont la taille est la moitié de celle de la population à risque. Dans l'étape suivante, on associera chaque ménage de cet échantillon à un autre ménage en dehors de l'échantillon, de telle sorte qu'à la fin toute la population à risque soit perturbée. L'échantillon est stratifié selon le nombre d'observations par maille du niveau géographique le plus fin, avec une probabilité d'être tiré proportionnelle à la moyenne arithmétique de deux indicateurs (jugés comme étant de bons prédicteurs de la divulgation) :

— un premier qui augmente avec la part de ménages à risque dans la zone (cette proportion est connue dans la population totale par construction) ;
— un second qui diminue avec le nombre de ménages dans la zone.

Tous les ménages ont ainsi une probabilité non nulle d'être dans l'échantillon, mais les ménages à risque ont une probabilité plus élevée. De plus, l'échantillon contient toujours au moins un ménage par zone géographique. L'algorithme, tel qu'il est proposé, permet aussi de plafonner la part de ménages d'une zone appartenant à l'échantillon, même si les résultats présentés ci-après n'ont pas utilisé cette possibilité.

Étape 3 : *Matching*

Le principe de l'étape de *matching* (ou appariement) est de trouver, pour chaque ménage de l'échantillon, une correspondance en dehors de l'échantillon, mais qui présente des caractéristiques géographiques et/ou démographiques proches, en privilégiant les autres ménages à risque. Tel que le propose l'algorithme, le processus d'appariement comprend différentes étapes et sous-étapes. Les observations sont traitées par ordre décroissant de leur niveau géographique de risque (niveau 3 en premier, puis 2, puis 1). Pour chacune de ces trois étapes, les contraintes de similarité sont de moins en moins strictes au fur et à mesure des sous-étapes.

12. Nous remercions Keith Spicer et Peter Youens pour leurs précieux conseils et éclaircissements au sujet de leur algorithme.

Plus précisément, si l'étape en cours consiste à traiter les ménages à risque de niveau l, le principe de chaque sous-étape est le suivant : on isole les ménages de l'échantillon à risque pour le niveau l, et l'on recherche pour chacun, une sorte de "jumeau" dans une "réserve". Ce jumeau est recherché aléatoirement en dehors de l'échantillon, mais doit satisfaire les trois conditions suivantes :

1. avoir le même "profil" ;
2. faire partie d'une autre zone géographique (de niveau l) ;
3. se trouver au sein de la même zone géographique au niveau hiérarchique supérieur (niveau $l+1$) [13]. Par exemple, pour un ménage à risque avec un niveau géographique de risque 1, le ménage jumeau sera recherché en dehors de la même zone de niveau 1 mais à l'intérieur de la même zone de niveau 2.

Les autres ménages à risque sont privilégiés. À la fin de chaque étape, l'échantillon et la réserve diminuent tous deux du nombre de ménages appariés, de sorte qu'il soit impossible d'échanger plusieurs fois un même ménage. Au fur et à mesure des sous-étapes, on relâche progressivement la contrainte de similarité des profils, de sorte qu'à la fin, tous les ménages de l'échantillon aient été appariés avec un autre ménage. Cette démarche garantit que presque tous les ménages à risque subissent une perturbation.

Étape 4 : *Swapping*

Enfin, les informations géographiques sont permutées entre les ménages appariés. La méthode n'introduit pas de "faux zéros" dans les décomptes d'individus par maille (puisqu'un ménage est toujours remplacé par un autre), mais peut en introduire dans les fréquences de variables.

14.3.2 Données et paramètres

Données

Pour ce chapitre, le TRS a été testé sur des données fiscales exhaustives [14]. Malheureusement, ces tests n'ont été effectués que sur la région Corse (la plus petite région française), car cela permettait de tester de nombreux paramètres dans un temps de calcul raisonnable. Ces tests ont été menés à des fins expérimentales ; pour généraliser les conclusions il faudrait les étendre à d'autres régions plus peuplées.

Pour cet algorithme de TRS, chaque unité des 3 niveaux géographiques nécessaires doit contenir un nombre d'observations minimum. Nous avons donc créé des unités géographiques *ad hoc* avec l'algorithme d'agrégation géographique de l'Insee présenté en section 14.2.2. Des carreaux de 1 km^2 sont regroupés pour constituer des rectangles (voir figure 14.4), ce qui aboutit à la structure hiérarchique suivante :

— niveau 3 : NUTS 3 (départements) ;
— niveau 2 : "gros rectangles" contenant au moins 5 000 individus, intersectés avec le niveau 3 ;
— niveau 1 : "petits rectangles" contenant au moins 100 individus, imbriqués dans les rectangles de niveau 2 (voir figure 14.5) et intersectés avec le niveau 3.

Chaque niveau est obtenu en désagrégeant le niveau précédent, et la maille la plus petite est le carreau de 1 km^2. Si un carreau de 1 km^2 contient au moins 100 individus, il peut constituer à lui seul une unité de niveau 1.

13. Plus précisément encore, l'algorithme procède à un ensemble d'itérations. Pour chaque itération, on rapproche aléatoirement chaque ménage à un autre ménage potentiel, et la paire est validée si toutes les conditions susmentionnées sont remplies. Si tel est le cas, les deux ménages sortent respectivement de l'échantillon et de la réserve. Sinon, ils y restent pour l'itération suivante. Un nombre maximum d'itérations est déterminé, suffisamment élevé pour qu'à la fin, plus aucun appariement possible ne soit détecté.

14. Nous n'avons pas testé la méthode sur les données du recensement français car il a la spécificité d'être une enquête avec un système de pondérations associé. Or, ce chapitre n'a pas pour objectif de traiter les questions de pondération.

La Corse compte 2 976 carrés de 1 km² mais uniquement 756 petits rectangles (qui contiennent au moins 100 habitants) et 39 gros rectangles (qui contiennent au moins 5 000 habitants, figure 14.5). Même si les rectangles sont créés pour atteindre un seuil d'observations (5 000 ou 100), tous les rectangles d'un même niveau n'ont pas le même nombre de ménages, puisque certains carreaux de 1 km² contiennent plus de 5 000 ménages dans les grandes villes (Ajaccio ou Bastia).

Dans ces tests, l'objectif n'est pas de diffuser les données sur des carreaux, mais sur des groupes de carreaux qui correspondent au niveau 1. Si l'objectif était de diffuser des comptages à un niveau plus fin (carreau de 1 km² par exemple), des clés de répartition devraient être définies, par exemple aléatoires dans les carreaux habités de la zone de niveau 1, ou proportionnellement au nombre d'habitants du carreau si cette quantité est connue (non jugée sensible).

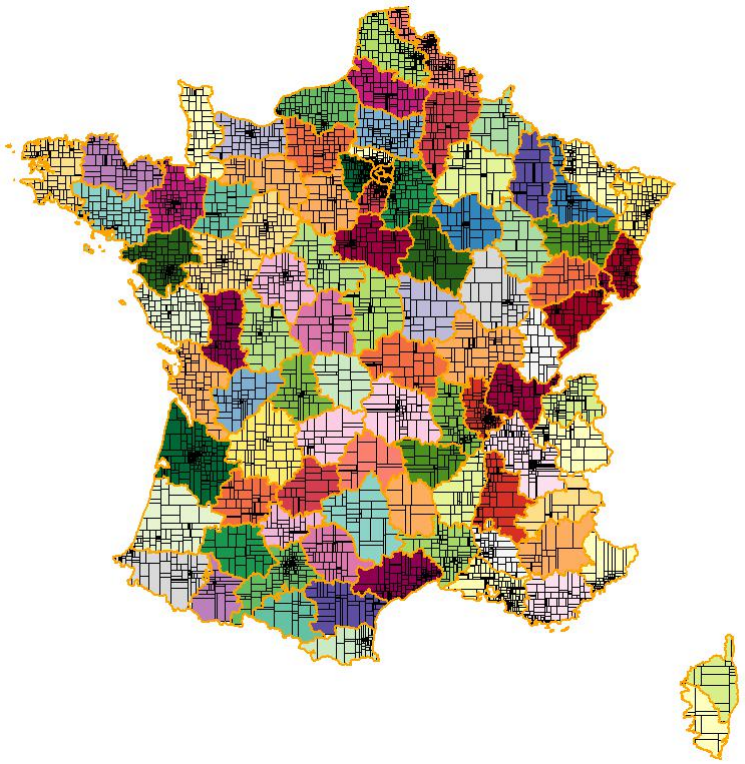


FIGURE 14.4 – La France découpée en rectangles de 5 000 individus (construits à partir du niveau NUTS 3)
Source : *Insee, Fideli 2015*

Paramètres

Tout d'abord, nous avons choisi 4 variables catégorielles pour définir le risque de divulgation : genre, tranche d'âge quinquennale, lieu de naissance (12 modalités) et lieu de résidence de l'année précédente (7 modalités).

Ensuite, le paramètre majeur de la méthode est le seuil en dessous duquel on considère qu'une observation est à risque. La taille de l'échantillon, et donc la part de ménages échangés, en découlent, même s'il n'y a pas de formule directe entre les deux. Par construction, la part d'individus échangés dans la population sera légèrement supérieure à ce paramètre, mais du même ordre de grandeur. Différentes valeurs (centiles d'ordre 1 à 10) ont été testées pour ce paramètre, induisant des parts d'individus échangés entre 2 % et 16 % [15].

15. Pour une autre région plus peuplée, la part d'individus échangés serait plus proche du paramètre initial. La

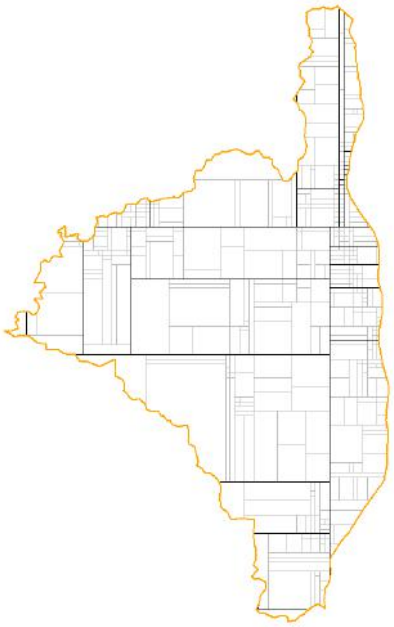


FIGURE 14.5 – Département 2B (Haute-Corse) découpé en rectangles de niveaux 1 et 2
Source : *Insee, Fideli 2015*

Enfin, 3 profils sont définis, du moins détaillé au plus détaillé. Deux ménages ne seront pas échangés s'ils ne partagent pas le même profil. Pour les tests qui suivent, nous avons choisi :

— profil A (le plus détaillé) : même nombre d'individus pour 7 catégories genre×âge [16] ;
— profil B (intermédiaire) : même nombre d'individus pour 5 catégories genre×âge (plus regroupées qu'en profil A) ;
— profil C (le moins détaillé) : même nombre d'individus dans le ménage.

14.3.3 Résultats

L'algorithme produit en sortie un fichier de microdonnées perturbé contenant, pour chaque ménage, sa localisation avant et après TRS. Les comptages sont effectués sur ce fichier.

On résume les résultats à travers une analyse risque-utilité (voir section 14.2.3). On évalue le risque par le seuil, paramètre majeur de la méthode (de 1 à 10 %) : un seuil élevé signifie que l'on accepte un faible niveau de risque [17].

Pour évaluer la perte d'utilité, on choisit les indicateurs suivants [18] :

— part des unités de niveau 1 (petits rectangles) perturbés (c'est-à-dire dont les comptages ne sont pas identiques), pour deux variables : nombre d'hommes (faisant partie des variables de *matching*) et nombre de personnes nées en France (n'en faisant pas partie) ;
— moyenne des valeurs absolues des écarts entre les comptages avant et après perturbation, en pourcentage de la valeur initiale (appelée ci-dessous AAD pour *Average Absolute Deviation*),

raison de cela est que, dans le cas de la Corse, les contraintes de *matching* sont plus difficiles à satisfaire dans le voisinage, et le jumeau se trouve souvent en dehors de la population à risque.

16. Il y a un nombre impair de catégories car pour certaines catégories d'âge, les hommes et les femmes sont regroupés.

17. Nous avons également envisagé une autre évaluation du risque, avec le 90e centile du score de rareté, défini comme vu précédemment (moyenne des inverses des fréquences par unité géographique de niveau 1), mais cet indicateur ne variait pas suffisamment pour être parlant.

18. La part d'individus échangés n'en fait pas partie puisque dans cette méthode, elle est fortement liée au seuil-paramètre par construction.

pour les comptages dans les unités de niveau 1 (petits rectangles), et pour ces deux mêmes variables ;

— indice de Moran, calculé au niveau des petits rectangles, pour 4 variables présentant une forte autocorrélation spatiale et pouvant être considérées comme sensibles : le nombre de personnes nées en France, le nombre d'enfants de moins de 5 ans, le revenu, et le nombre de personnes vivant dans un quartier de la politique de la ville (QPV).

Les résultats sont consignés dans la table 14.1 et la figure 14.6 (*R-U maps* légèrement différentes que ce qui avait été suggéré plus haut). Les variables auxquelles on s'intéresse, soit sont directement prises en compte dans la méthode *via* le profil de *matching* ($V1$, nombre d'hommes), soit sont indirectement prises en compte *via* le calcul du score de rareté ($V2$, nombre de personnes nées en France ou $V3$, nombre d'enfants de moins de 5 ans), soit ne le sont pas du tout ($V4$, revenu ou $V5$, nombre de personnes en QPV).

Évaluation du risque (%)									
Seuil (paramètre de la méthode)	0	1	2	3	4	5	7	8	10
Évaluation de la perte d'utilité (%)									
Part d'individus perturbés	0	2	4	5	7	8	11	13	16
Part d'unités de niveau 1 perturbées – V1	0	38	52	56	63	69	73	75	78
Part d'unités de niveau 1 perturbées - V2	0	71	82	85	85	88	90	92	94
AAD (niveau 1) - V1	0,0	0,5	0,8	0,9	1,1	1,2	1,5	1,6	1,7
AAD (niveau 1) - V2	0,0	0,8	1,1	1,3	1,6	1,6	2,1	2,2	2,5
Indice de Moran (niveau 1) : V2	6,5	6,5	6,5	6,5	6,5	6,5	6,5	6,5	6,5
Indice de Moran (niveau 1) : V3	6,5	6,4	6,5	6,5	6,6	6,7	6,7	6,8	6,4
Indice de Moran (niveau 1) : V4	5,5	5,2	5,1	4,6	5,2	6,8	8,2	5,7	6,4
Indice de Moran (niveau 1) : V5	7,7	7,7	7,8	7,7	7,8	7,7	7,3	7,2	7,3

V1 : nombre d'hommes
V2 : nombre de personnes nées en France
V3 : nombre d'enfants de moins de 5 ans
V4 : revenu moyen
V5 : nombre de personnes en QPV

TABLE 14.1 – Résultats des tests de la méthode TRS menés sur la Corse pour plusieurs paramètres
Source : *Insee, Fideli 2015*
Note : Pour un niveau de risque de 1 % (autrement dit si l'on considère que les 1 % d'individus les plus rares sont à risque), la méthode testée conduit à permuter entre eux 2 % des individus. 38 % des petits rectangles voient leur total modifié pour la variable V1. En moyenne, un petit rectangle voit sa valeur initiale modifiée de +/-0.5 % pour V1. L'indice de Moran de la variable V2 calculé au niveau des petits rectangles (voisinage de la reine) est inchangé par rapport à une absence de perturbation (6,5 % comme avec le risque de 0 %)

Plus le niveau de risque acceptable est élevé (autrement dit plus la part de population à risque est petite), plus la perturbation est faible (part de petits rectangles perturbés ou écart absolu moyen).

Même pour des petites valeurs de seuil, la majorité des petits rectangles sont perturbés (pour le paramètre 1 %, soit le niveau de risque testé le plus élevé, 70 % des petits rectangles sont perturbés pour la variable "nombre de personnes nées en France"). Le niveau de perturbation est cependant

raisonnable : pour le niveau de risque le plus élevé (paramètre égal à 1 %), 0,4 % et 0,7 % (pour $V1$ et $V2$, respectivement) des écarts absolus sont inférieurs à 5 % du décompte et l'AAD est inférieur à 1 %. Pour le plus faible niveau de risque testé (paramètre égal à 10 %), l'AAD est de 2,5 % pour le nombre de personnes nées en France et de 1,7 % pour le nombre d'hommes.

On s'intéresse maintenant aux corrélations spatiales, en regardant dans quelle mesure l'indicateur de Moran (calculé pour les unités de niveau 1) est modifié avant et après TRS. On constate alors que cette distorsion peut être très importante (jusqu'à 50 % de variation de l'indicateur), qu'elle n'est pas toujours dans le même sens (le niveau d'autocorrélation spatiale peut augmenter ou diminuer après application du TRS), et qu'elle n'est pas une fonction monotone du niveau de risque.

Enfin, on observe également que la perte d'utilité, quel que soit l'indicateur considéré pour l'appréhender, varie beaucoup selon la variable d'intérêt. Pour la variable directement prise en compte dans la méthode (*via* la définition du profil : $V1$ dans les tests), la perturbation est moindre que pour les variables indirectement prises en compte (*via* l'identification des individus à risque : $V2$ et $V3$ dans les tests), et *a fortiori* que pour les variables qui n'interviennent pas du tout dans la méthode ($V4$ ou $V5$ dans les tests).

Coefficient de Pearson	V1	V2	V3	V4	V5	V6
V1 (nombre d'hommes)	1	1,00	0,97	0,13	0,97	*1,00*
V2 (nombre de personnes nées en France)	1,00	1	0,96	0,14	0,97	*1,00*
V3 (nombre d'enfants de moins de 5 ans)	0,97	0,96	1	0,14	0,93	*0,97*
V4 (revenu moyen)	0,13	0,14	0,14	1	0,15	0,13
V5 (nombre de personnes en QPV)	0,97	0,97	0,93	0,15	1	*0,97*
V6 (nombre total de personnes)	*1,00*	*1,00*	*0,97*	*0,13*	*0,97*	*1*

Remarque : V1, V2, V3 et V5 sont des totaux donc sont fortement corrélés au nombre total de personnes dans le rectangle, tandis que V4 est une moyenne.

TABLE 14.2 – Coefficients de Pearson entre les différentes variables d'intérêt
Source : *Insee, Fideli 2015*

Plus précisément sur la déformation des corrélations spatiales : le I de Moran est inchangé pour la variable qui définit le profil d'appariement ($V1$), et il ne varie que légèrement pour les variables indirectement prises en compte ($V2$ et $V3$). Il est également légèrement modifié pour les variables fortement corrélées avec le profil de *matching* ($V5$, voir figure 14.2). *A contrario*, les corrélations spatiales peuvent être très déformées pour les variables qui ne sont pas corrélées avec le profil de *matching* ($V4$).

La déformation du I de Moran n'augmente pas particulièrement avec le niveau de risque, mais des comportements erratiques peuvent apparaître, du fait du caractère aléatoire de l'algorithme pendant l'étape de *matching*. Puisque la méthode ne considère pas le revenu ($V4$) comme une variable à conserver et n'en fait pas une variable de *matching*, et puisque cette variable n'est pas corrélée avec une autre variable que l'on souhaite conserver, les ménages dont les revenus sont similaires peuvent se trouver soit rapprochés soit éloignés, d'une exécution de la méthode à l'autre.

14.4 Problèmes de différenciation géographique

14.4.1 Définition

La différenciation géographique se produit lorsqu'un intrus est en mesure de combiner des données diffusées dans différentes géographies afin de reconstituer des statistiques sur une zone plus

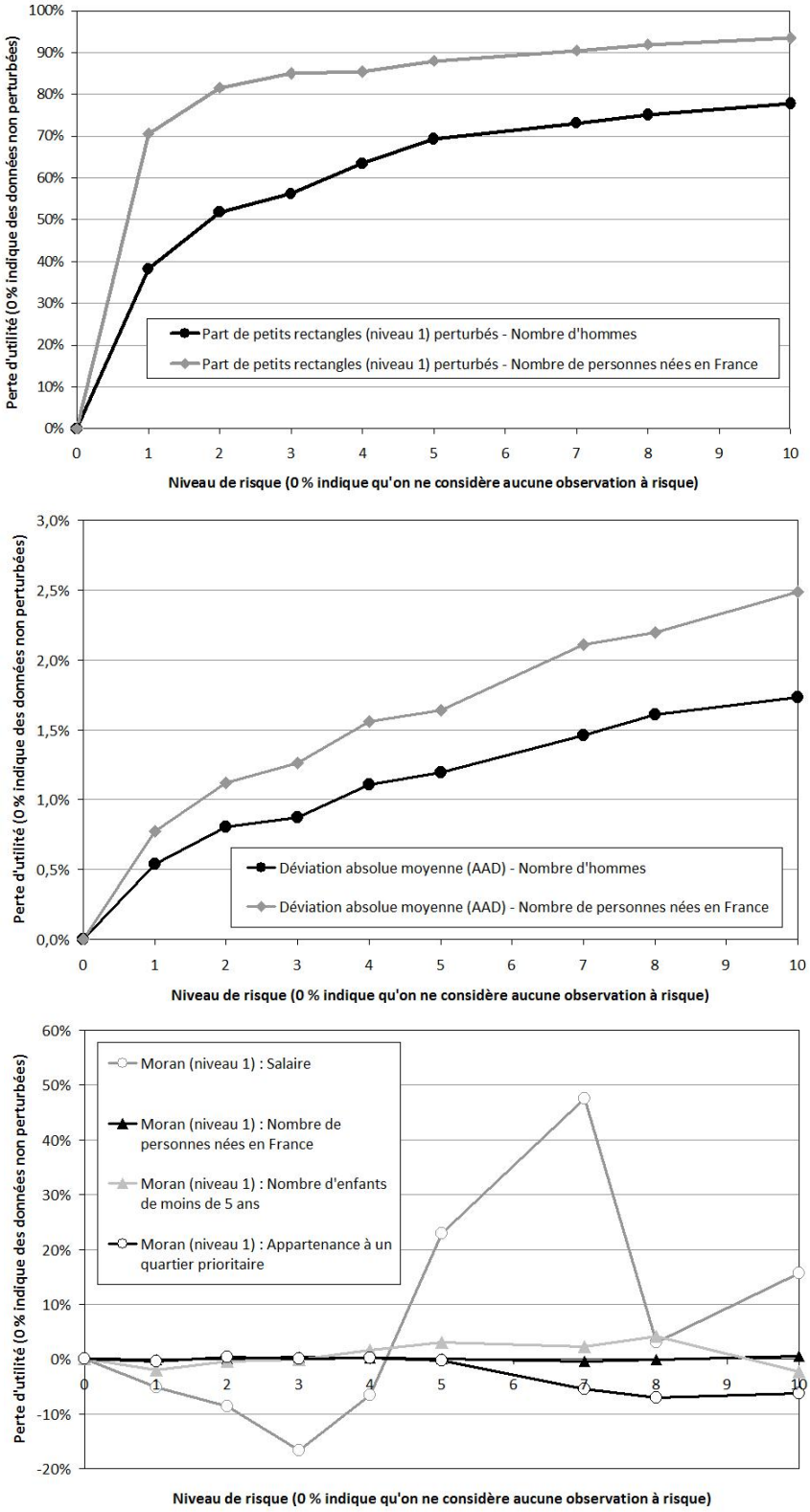


FIGURE 14.6 – Perte d'utilité en fonction du niveau de risque, pour 3 indicateurs de perte d'utilité
Source : *Insee, Fideli 2015*

petite ou de déduire le lieu auquel une observation se rapporte. Ce problème a été présenté dans de nombreux articles. Il est souvent évoqué au sujet des données de recensement (DUKE-WILLIAMS et al. 1998, ONU 2004), mais se pose pour la diffusion de n'importe quelle source.

Dans les systèmes de géographies imbriquées (par ex. régions – départements – villes), le problème est assez facile à résoudre car les "petites" zones pouvant être déduites par soustraction sont directement liées à la hiérarchie des différentes géographies. Une fois que l'ensemble des petites zones à protéger est identifié (ce que l'on appelle secret primaire), on utilise en général un logiciel de confidentialité comme Tau-Argus pour gérer le secret secondaire, à savoir l'ensemble des zones à traiter pour qu'il soit impossible de reconstituer les données du secret primaire. Dans un système de géographies imbriquées représenté par un arbre hiérarchique, le problème est le même qu'avec toutes les autres variables d'intérêt hiérarchisées utilisées dans une nomenclature (par ex. la suite sections - divisions - groupes - classes utilisée dans la classification NACE des activités économiques).

Le problème de la différenciation géographique devient plus complexe lorsque les différentes géographies utilisées pour la diffusion ne sont pas imbriquées (ABS 2015). Dans ce cas, il n'y a aucun arbre hiérarchique sous-jacent et des algorithmes spécifiques doivent être mis en œuvre pour identifier toutes les soustractions auxquelles un intrus peut procéder entre les différentes zones pour déduire des statistiques sur des zones plus petites.

Le problème de différenciation s'aggrave lorsque la taille de la maille de diffusion diminue (typiquement dans le cas de données carroyées). Il augmente également avec le nombre de géographies, et d'autant plus lorsque celles-ci ne sont pas hiérarchiques : par exemple, quand des INS diffusent des statistiques sur un zonage *ad hoc* dans le cadre d'un partenariat spécifique ou lorsque l'utilisateur peut dessiner une zone à façon *via* un web-service dédié.

Le problème de différenciation intervient également quand des mêmes statistiques sont diffusées pour des dates différentes. Par exemple, dans le cas de statistiques d'entreprises publiées chaque année, l'intrus pourrait rapprocher les différentes versions pour tenter de trouver des valeurs cachées. La méthode SDC choisie pour une diffusion devrait alors tenir compte des choix qui ont été faits et des valeurs cachées lors les diffusions précédentes.

14.4.2 Illustration

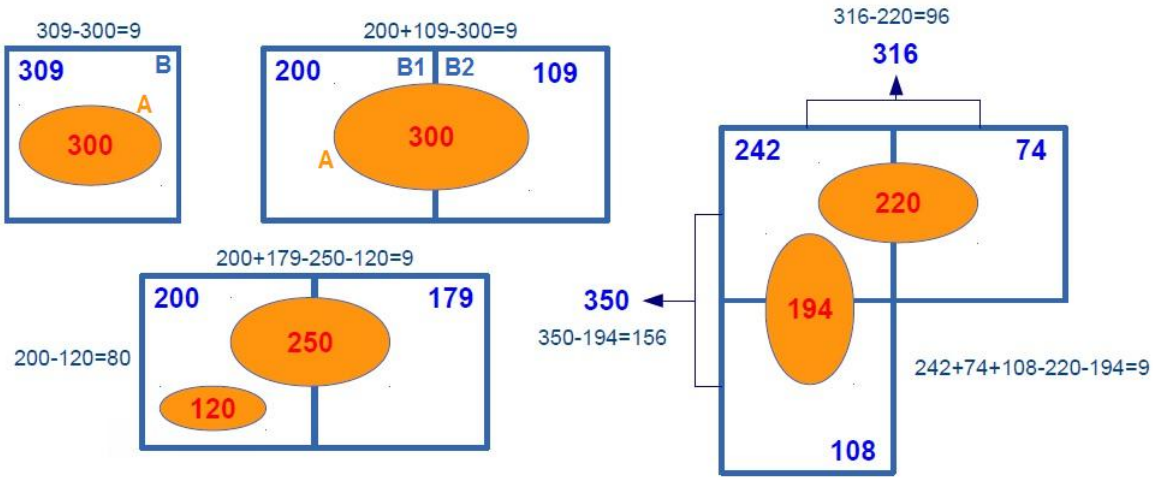


FIGURE 14.7 – Exemples de rupture de confidentialité par différenciation géographique

La figure 14.7 présente des cas possibles de différenciation géographique. Les zones qui se chevauchent entre les ovales (A) et les rectangles (B) sont colorées en orange. Dans le premier

cas, le zonage B englobe le zonage A : l'intrus peut reconstruire les informations au sujet de B-A par soustraction, et divulguer de données d'un petit nombre d'individus (9). Dans la deuxième configuration, l'intrus peut combiner deux zones du zonage B pour effectuer l'opération (B1 + B2) - A et ainsi déduire le comptage d'une zone non diffusée. Les deux dernières configurations de la figure 14.7 montrent que la différenciation est possible avec n'importe quelle combinaison des deux zonages.

Dans ces mêmes exemples, en supposant que l'information ne peut être diffusée si elle concerne moins de 10 individus, alors il y a rupture du secret par différenciation géographique dans chacune de ces 4 configurations. Dans les autres cas de chevauchement, l'intrus ne peut pas déduire directement d'information sur une nouvelle zone, mais un problème peut subsister si l'intrus se met à mobiliser de l'information auxiliaire sur la zone englobant la zone de chevauchement. Il faut également prendre en compte les éléments naturels de la zone, puisqu'il est impossible par exemple que des individus résident dans un lac ou sur une autoroute. Ces zones inhabitées ne doivent pas intervenir pour protéger d'autres données et doivent être diffusées en tant que telles.

14.4.3 Identification des zones à risque

Pour résoudre le problème de différenciation, la première étape est d'identifier les zones exposées au risque de divulgation. Cela peut être relativement simple grâce aux systèmes d'information géographique (SIG), mais devient lourd à mesure que le nombre de géographies non imbriquées augmente, puisqu'il s'agit d'un problème NP-difficile.

L'identification des zones à risque a lieu en deux étapes : identification des zones touchées par le secret primaire, puis celles touchées par le secret secondaire. Un processus classique consiste à rechercher tous les chevauchements possibles entre les géographies non imbriquées. Comme l'illustre la figure 14.7, des problèmes peuvent apparaître en combinant plusieurs mailles d'un même zonage. Un critère de confidentialité doit être déterminé, par exemple au moins 10 individus par zone.

Si l'une des géographies non imbriquées est hiérarchique, il faut veiller à vérifier que les totaux soient conservés (par exemple lorsque l'on considère deux zonages : d'un côté, la géographie imbriquée région - département - ville et, de l'autre, un zonage spécifique réalisé pour un partenaire).

Il semble important de limiter la perte d'information, et donc d'inclure des règles d'optimisation qui limitent le nombre d'individus supprimés, ou priorisent la diffusion des zones sur lesquelles on souhaite à tout prix que l'information soit diffusée, par exemple les quartiers prioritaires. Cette étape d'exploration de tous les chevauchements possibles requiert une grande puissance de calcul.

14.4.4 Méthodes de protection

Différentes méthodes peuvent permettre de rétablir la confidentialité en présence de problèmes de différenciation dus à des chevauchements.

— Première possibilité, le zonage peut être modifié : les frontières peuvent être déplacées de manière à neutraliser les zones de chevauchement, par exemple en imbriquant les différentes géographies et en créant un arbre hiérarchique strict.

— Deuxième possibilité, si les frontières sont fixes, on peut procéder à des fusions de zones pour éliminer les chevauchements. Cela réduit le niveau de détail mais permet une diffusion exhaustive.

— Une troisième méthode consiste à supprimer les données sur des zones de chevauchement. À cause des contraintes de diffusion, cette option est souvent choisie, car elle propose un compromis entre un niveau de détail acceptable et une faible part d'observations supprimées.

— Plutôt que de supprimer des données lorsque les frontières et le zonage sont fixes, une quatrième option est de perturber les données, par exemple en ajoutant du bruit aux comptages des zones à risque. Pour y procéder de manière consistante entre plusieurs tableaux ou plu-

sieurs géographies, l'*Australian Bureau of Statistics* (ABS) a élaboré une technique appelée *cell-key method*, qui attribue une "clé" à chaque observation du fichier de microdonnées, et l'utilise pour conserver la cohérence entre les tableaux diffusés sur différentes géographies (FRASER et al. 2005). La *cell-key method* a été adaptée par l'ONS pour diffuser des données du recensement britannique et a également été testée dans le cadre du programme d'Eurostat "Protection harmonisée des données de recensement du SSE".

Conclusion

La discussion globale sur les méthodes de confidentialité va de pair avec une réflexion plus stratégique sur ce que veulent vraiment diffuser les INS *in fine*. Parmi les éléments cruciaux de cette réflexion : peut-on assumer la diffusion d'informations perturbées ? À quel point craint-on les erreurs d'interprétation d'un utilisateur trop pressé ? Les choix méthodologiques doivent être faits, dans la mesure du possible, en concertation avec les utilisateurs potentiels, dans l'optique de ne pas détériorer les analyses futures.

Gérer la confidentialité de données spatiales peut être vu comme une opportunité pour affiner les méthodes SDC, puisque la densité de population et la similitude avec ses voisins sont des prédicteurs fondamentaux du risque de divulgation. Dans l'état de l'art actuel, l'information géographique est prise en compte par des méthodes prétabulées qui utilisent l'information du voisinage (imputation locale, *targeted record swapping*). Une prise en compte plus fine des coordonnées géographiques en évaluant plus localement la densité pourrait être envisagée à l'avenir, au gré de l'augmentation des capacités de calcul. Cependant, les gains de précision se font au prix d'un surcroît de complexité de la méthode de protection, et donc de difficultés supplémentaires pour communiquer pédagogiquement à son sujet auprès des utilisateurs.

Des tests, menés sur les données fiscales exhaustives d'une région française, ont démontré que pour des niveaux de risque raisonnables, la méthode de *targeted record swapping* implique une bonne conservation des corrélations spatiales, même si ces tests mériteraient d'être poursuivis avec d'autres méthodes et sur des régions plus peuplées. Malgré ces bons résultats, il apparaît qu'appliquer uniquement une méthode prétabulée ne suffit pas, d'une part car elle demanderait de perturber trop d'observations pour atteindre un niveau de risque global acceptable, et d'autre part à cause de la perception du public. Les méthodes post-tabulées, elles, font plus clairement apparaître la protection contre la divulgation. C'est pourquoi, dans le cadre de la diffusion du recensement, Eurostat conseille aux INS de combiner des méthodes prétabulées prenant en compte l'information géographique, et des méthodes post-tabulées.

Quelle que soit la méthode utilisée, et quand bien même il s'agit de la plus basique, il est intéressant d'évaluer dans quelle mesure elle dégrade les relations spatiales de certains attributs. À cette fin, on peut tracer des cartes risque-utilité, en choisissant les indicateurs d'autocorrélation spatiale comme métrique de perte d'utilité. Même si les paramètres précis utilisés doivent rester cachés aux utilisateurs, il est essentiel que les INS documentent la méthode mise en œuvre et les choix effectués. C'est la condition pour que les utilisateurs aient conscience que les données analysées ont été perturbées ou peuvent ne pas être exhaustives. Par exemple, l'expert en confidentialité peut informer les utilisateurs potentiels de la déformation de l'indice de Moran ou des indicateurs LISA induite par la méthode de protection, pour les prévenir d'analyses fallacieuses.

Références - Chapitre 14

ABS (2015). « SSF Guidance Material – Protecting Privacy for Geospatially Enabled Statistics : Geographic Differencing ».

ARMSTRONG, Marc P, Gerard RUSHTON, Dale L ZIMMERMAN et al. (1999). « Geographically masking health data to preserve confidentiality ». *Statistics in medicine* 18.5, p. 497–525.

BACKER, Lars H et al. (2011). « GEOSTAT 1A : Representing Census data in a European population grid ». *Final Report*.

BEHNISCH, Martin et al. (2013). « Using Quadtree representations in building stock visualization and analysis ». *Erdkunde*, p. 151–166.

BERGEAT, Maxime (2016). « La gestion de la confidentialité pour les données individuelles ». *Document de travail INSEE M2016/07*.

BROWN, D (2003). « Different approaches to disclosure control problems associated with geography ». *Proceeding of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*.

CLARKE, John (1995). « Population and the environment : complex interrelationships. »

CLIFTON, Kelly et Nebahat NOYAN (2012). « Framework for Applying Data Masking and Geo-Perturbation Methods to Household Travel Survey Datasets ». *91st Annual Meeting of Transportation Research Board, Washington, DC*.

CURTIS, Andrew J, Jacqueline W MILLS et Michael LEITNER (2006). « Spatial confidentiality and GIS : re-engineering mortality locations from published maps about Hurricane Katrina ». *International Journal of Health Geographics* 5.1, p. 44–55.

DE WOLF, PP (2015). « Public use files of eu-silc and eu-lfs data ». *Joint UNECE-Eurostat work session on statistical data confidentiality, Helsinki, Finland*.

DEICHMANN, Uwe, Deborah BALK et Greg YETMAN (2001). « Transforming population data for interdisciplinary usages : from census to grid ». *Washington (DC) : Center for International Earth Science Information Network* 200.1.

DOMINGO-FERRER, Josep, Josep M MATEO-SANZ et Vicenç TORRA (2001). « Comparing SDC methods for microdata on the basis of information loss and disclosure risk ». *Pre-proceedings of ETK-NTTS*. T. 2, p. 807–826.

DOMINGO-FERRER, Josep et Rolando TRUJILLO-RASUA (2011). « Anonymization of trajectory data ».

DOYLE, Pat et al. (2001). « Confidentiality, disclosure, and data acces : theory and practical applications for statistical agencies ».

DUKE-WILLIAMS, Oliver et Philip REES (1998). « Can Census Offices publish statistics for more than one small area geography ? An analysis of the differencing problem in statistical disclosure ». *International Journal of Geographical Information Science* 12.6, p. 579–605.

DUNCAN, George T, Sallie A KELLER-MCNULTY et S Lynne STOKES (2001). « Disclosure risk vs. data utility : The RU confidentiality map ». *Chance*. Citeseer.

DUNCAN, George T et Diane LAMBERT (1986). « Disclosure-limited data dissemination ». *Journal of the American statistical association* 81.393, p. 10–18.

ELLIOT, Mark J et al. (2005). « SUDA : A program for detecting special uniques ». *Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, p. 353–362.

ELLIOT, Mark et Josep DOMINGO-FERRER (2014). « EUL to OGD : A simulated attack on two social survey datasets ». *Privacy in Statistical Databases*. Sous la dir. de Josep DOMINGO-FERRER.

FRASER, Bruce et Janice WOOTON (2005). « A proposed method for confidentialising tabular output to protect against differencing ». *Monographs of Official Statistics. Work session on Statistical Data Confidentiality*, p. 299–302.

GOMATAM, Shanti et al. (2005). « Data dissemination and disclosure limitation in a world without microdata : A risk-utility framework for remote access analysis servers ». *Statistical Science*, p. 163–177.

GOUWELEEUW, JM, Peter KOOIMAN et PP DE WOLF (1998). « Post randomisation for statistical disclosure control : Theory and implementation ». *Journal of official Statistics* 14.4, p. 463–478.

HALDORSON, Marie et al. (2017). « A Point-based Foundation for Statistics : Final report from the GEOSTAT 2 project ». *Final Report*.

HETTIARACHCHI, Raja (2013). « Data confidentiality, residual disclosure and risk mitigation ». Working Paper for joint UNECE/Eurostat Work Session on Statistical Data Confidentiality.

HUNDEPOOL, Anco et al. (2010). « Handbook on statistical disclosure control ». *ESSnet on Statistical Disclosure Control*.

HUNDEPOOL, Anco et al. (2012). « Statistical disclosure control ».

INSEE (2010). « Guide du secret statistique ». *Documentation INSEE*.

ITO, Shinsuke et Naomi HOSHINO (2014). « Data swapping as a more efficient tool to create anonymized census microdata in Japan ». *Privacy in Statistical Databases*, p. 1–14.

KAMLET, MS, S KLEPPER et RG FRANK (1985). « Mixing micro and macro data : Statistical issues and implication for data collection and reporting ». *Proceedings of the 1985 Public Health Conference on Records and Statistics*.

LAMBERT, Diane (1993). « Measures of disclosure risk and harm ». *Journal of Official Statistics* 9.2, p. 313–331.

LONGHURST, Jane et al. (2007). « Statistical disclosure control for the 2011 UK census ». *Joint UNECE/Eurostat conference on Statistical Disclosure Control, Manchester*, p. 17–19.

MARKKULA, Jouni (1999). « Statistical disclosure control of small area statistics using local restricted imputation ». *Bulletin of the International Statistical Institute (52nd Session)*, p. 267–268.

MASSELL, Paul, Laura ZAYATZ et Jeremy FUNK (2006). « Protecting the confidentiality of survey tabular data by adding noise to the underlying microdata : Application to the commodity flow survey ». *Privacy in Statistical Databases*. Springer, p. 304–317.

NAGY, Beata (2015). « Targeted record swapping on grid-based statistics in Hungary ». *Submission for the 2015 IAOS Prize for Young Statisticians*.

ONS (2006). « Review of the Dissemination of Health Statistics : Confidentiality Guidance ». *Working Paper 5 : References and other Guidance*.

ONU (2004). « Manuel des systèmes d'information géographique et de cartographie numérique ». F-79, p. 118–119.

SHLOMO, Natalie (2005). « Assessment of statistical disclosure control methods for the 2001 UK Census ». *Monographs of official statistics*, p. 141–152.

— (2007). « Statistical disclosure control methods for census frequency tables ». *International Statistical Review* 75.2, p. 199–217.

SHLOMO, Natalie et Jordi MARÉS (2013). « Comparison of Perturbation Approaches for Spatial Outliers in Microdata ». *the Cathie March Centre for Census and Survey Research*.

SHLOMO, Natalie, Caroline TUDOR et Paul GROOM (2010). « Data Swapping for Protecting Census Tables. » *Privacy in statistical databases*. Springer, p. 41–51.

TAMMILEHTO-LUODE, Marja (2011). « Opportunities and challenges of grid-based statistics ». *World Statistics Congress of the International Statistical Institute*.

TORRA, Vicenc et Michael CARLSON (2013). « On the Hellinger distance for measuring information loss in microdata ». *Joint UNECE/Eurostat work session on statistical data confidentiality, Ottawa, Canada, 28-30 October 2013*.

VANWEY, Leah K et al. (2005). « Confidentiality and spatially explicit data : Concerns and challenges ». *Proceedings of the National Academy of Sciences* 102.43, p. 15337–15342.

WILLENBORG, Leon et Ton DE WAAL (2012). *Elements of statistical disclosure control*. T. 155. Springer Science & Business Media.

YOUNG, Caroline, David MARTIN et Chris SKINNER (2009). « Geographically intelligent disclosure control for flexible aggregation of census data ». *International Journal of Geographical Information Science* 23.4, p. 457–482.

ZIMMERMAN, Dale L et Claire PAVLIK (2008). « Quantifying the effects of mask metadata disclosure and multiple releases on the confidentiality of geographically masked health data ». *Geographical Analysis* 40.1, p. 52–76.

Index

Insee **Méthodes**

Afin de tenir compte des évolutions récentes dans l'accès aux données géographiques et de la demande croissante d'information statistique finement localisée, l'Insee a entrepris une refonte de son système d'information géographique, notamment pour réaliser et exploiter le recensement de la population. Pour ce faire, l'Insee s'appuie sur les nombreuses initiatives internationales (ONU, Eurostat, etc.) visant à établir de bonnes pratiques de gestion coordonnée des informations statistiques et géographiques.

En complément, l'Institut a souhaité valoriser sa capacité à analyser ce type d'information. La division des méthodes et référentiels géographiques de l'Insee a ainsi proposé à Eurostat et au Forum Européen de Statistique et de Géographie (EFGS en anglais) la rédaction collaborative d'un manuel d'analyse spatiale.
Ce document s'inscrit dans le développement, en cours, d'un système d'information statistique, finement géoréférencé, figurant au nombre des ambitions stratégiques de l'Insee à l'horizon 2025 et destiné à mieux couvrir les besoins de la statistique publique française. Il décrit un ensemble de méthodes statistiques mobilisables pour traiter des données géolocalisées. De la description des données spatiales jusqu'à la gestion de leur confidentialité en passant par l'économétrie spatiale, l'échantillonnage ou le lissage spatial, le manuel couvre un large spectre de sujets.

Si les fondements théoriques occupent une part significative de l'ouvrage, une place très importante est réservée au développement d'exemples concrets assortis de programmes de traitement (écrits en R). Le manuel peut ainsi servir de support à des actions de formation initiale ou professionnelle.

ISSN : 1142-3080
ISBN : 978-2-11-139684-5

Insee
Mesurer pour comprendre