UNIVERSITE PARIS-DAUPHINE
ECOLE DOCTORALE DE DAUPHINE

N° attribué par la bibliothèque

Statistiques spatiales et Etude immobilière

THESE
présentée et soutenue publiquement par
Piyawan SRIKHUM

Le 12 novembre 2012

pour l'obtention du titre de
DOCTEUR EN SCIENCES DE GESTION
(Arrêté du 7 août 2006)

JURY

Directeur de recherche :	**Laurent BATSCH** Professeur à l'Université Paris-Dauphine
Rapporteurs :	**Edwin DEUTSCH** Professeur à l'Université de Technologie de Vienne
	Catherine REFAIT-ALEXANDRE Professeur à l'Université de Franche-Comté
Suffragants :	**Denis ALLARD** Directeur de recherche à l'INRA Avignon
	Arnaud SIMON Maître de conférences à l'Université Paris-Dauphine
	André TORRE Directeur de recherche à l'INRA

REMERCIEMENTS

Je tiens tout d'abord à exprimer mes remerciements au Pr. Laurent Batsch, mon directeur de thèse qui a accepté d'assurer l'encadrement de ce travail dans ce domaine si particulier de la Finance Immobilière. Son soutien, ses encouragements et sa confiance sont pour beaucoup dans cette thèse.

Je remercie le Pr. Edwin Deutsch et le Pr. Catherine Refait-Alexandre, du grand honneur qu'ils me font en acceptant de participer au jury en tant que rapporteurs, ainsi que M. Denis Allard, M. Arnaud Simon et M. André Torre d'avoir accepté d'être membres du jury de soutenance malgré leurs responsabilités et, pour certains, leur éloignement.

Une reconnaissance particulière est également adressée à M. Arnaud Simon qui m'accompagne depuis mes débuts en recherche, qui m'a donné des conseils dans mes moments de doute, qui m'a aidé à développer des capacités de chercheur et qui m'a toujours encouragé à présenter mes travaux lors de congrès.

Je souhaite également adresser mes remerciements au Pr. Edwin Deutsch qui n'a pas hésité à m'accueillir pour passer un semestre d'échange à l'Université de Vienne. Ses conseils donnés durant ce semestre d'échange m'ont bien aidé à préciser ma problématique de recherche. Je tiens aussi à exprimer ma reconnaissance pour ses remarques constructives et ses précieux conseils, dont j'ai pu bénéficier lors de la pré-soutenance de cette thèse.

Il est impossible de ne pas évoquer tout le bien que je pense du département économie finance assurance et banque du CNAM. Je remercie Bertrand Djembissi, Nathalie Oriol et Frédéric Loss pour leur accueil au sein du département, leur conseil et leur soutien.

Je tiens à remercier toute l'équipe de recherche du DRM-Finance, grâce à laquelle j'ai pu bénéficier d'expériences de recherche privilégiées, mais également d'un financement pour mes conférences. Cette thèse n'aurait pu voir le jour sans l'intégration à cette équipe. J'ai une pensée toute particulière pour Timothée, Dorra, Sarah et Salma. Leur amitié, leur soutien et surtout leur humour ont donné l'ambiance dans la salle des doctorants.

Je souhaite remercier particulièrement Arnaud, Bertrand, Nathalie, Ghislaine, Timothée, Eric, Thomas et Yona qui ont participé à la relecture du document de thèse.

J'adresse une pensée affectueuse à ma famille. A ma mère pour sa présence si rassurante, mes deux sœurs, Waraporn et Monthana, pour le soutien inconditionnel, et à mon père qui m'a poussé à avancer dans cette direction mais qui n'aura malheureusement pas pu voir cette thèse achevée. Je suis également très reconnaissant à tous les officiers du bureau du conseiller de l'éducation à Paris, qui ont permis la réalisation de cette thèse dans les meilleures conditions.

Enfin, je termine par tous mes amis thaïlandais, Natsuda, Jarukit, Cherra, Artra, Pongsak, Somphop, Baramee, Naruepon, Ekkarin, Noka et Supot, dont leurs soutiens, leurs coups de téléphone, leurs guitares et leurs repas ont été indispensable à mon épanouissement tout au long de cette aventure à l'étranger.

SOMMAIRE

INTRODUCTION GÉNÉRALE

"Buying real estate is not only the best way, the quickest way, the safest way, but the only way to become wealthy"

Marshall Field the founder of Marshall Field and Company

L'immobilier occupe une place de plus en plus prépondérante dans les choix d'investissement des épargnants. Afin de réaliser un investissement immobilier, une bonne évaluation du bien est une étape indispensable pour les investisseurs. Une question naturelle se pose alors : **le bien immobilier peut-il être évalué comme un simple produit financier ?**

En considérant les caractéristiques spécifiques d'un bien immobilier telles que son caractère indivisible, sa valeur unitaire très élevée, sa faible liquidité, sa grande hétérogénéité et son immobilité physique, on est conduit à distinguer ce type d'actif des autres produits financiers. Ses méthodes d'estimation doivent donc être aussi différentes de celles des autres actifs et doivent pouvoir prendre en compte les caractéristiques propres aux biens immobiliers.

Rosen (1974) fonde l'approche hédonique qui permet d'intégrer l'hétérogénéité des biens immobiliers. La méthode des prix hédonistes permet d'estimer le prix des différentes caractéristiques : le prix de marché consiste en la somme des prix implicites attachés aux caractéristiques du bien. Cependant, comme son nom l'indique, le bien immobilier ne peut pas être déplacé ; sa valeur dépend donc aussi partiellement de sa localisation. La méthode d'estimation de sa valeur immobilière se doit aussi de prendre en compte cette caractéristique spatiale. **Comment prendre en compte la caractéristique spatiale des données immobilières dans les modèles d'évaluation ? Faut-il améliorer le modèle hédonique ou aller vers de nouveau modèle ?**

Le modèle de prix hédoniste standard peut être amélioré en intégrant des caractéristiques spatiales comme variables explicatives du modèle. Mais, malgré le nombre important des variables locales que l'on peut rajouter, en général les régressions n'aboutissent pas à des résidus spatialement non corrélés. Afin de déterminer le modèle

qui permet d'analyser plus finement cette dépendance spatiale, il faut identifier précisément le mode d'influence de la caractéristique spatiale sur le prix immobilier.

Considérons le processus d'évaluation d'un bien par un particulier : pour déterminer la valeur de son bien, le propriétaire peut se renseigner, soit auprès de l'expert de quartier qui donne une estimation de prix basée sur la valeur de transaction des biens voisins, soit directement auprès des propriétaires des biens proches. Ce processus d'évaluation crée localement une dépendance spatiale des prix. De plus des biens voisins ont souvent été construits à la même période, ils ont fréquemment la même structure, le même style et la même taille. Par ailleurs, ces biens doivent faire face aux mêmes variables d'externalité. Cette ressemblance locale crée donc un problème de corrélation spatiale des variables explicatives du modèle des prix hédonistes. Si cette dépendance spatiale n'est pas prise en considération lors de la spécification du modèle, les résidus du modèle hédonique seront dépendants. L'interaction spatiale entre les prix immobiliers est un phénomène complexe qui possède plusieurs dimensions ; il existe plusieurs sources de dépendance spatiale. Le modèle hédonique extensif ne peut pas corriger à lui seul cette dépendance. Le recours aux méthodes plus sophistiquées de la statistique spatiale est alors requis.

Antécédemment, la statistique spatiale a été appliquée dans les sciences de l'environnement, les sciences de la terre, en épidémiologie, en agronomie, en météorologie, en géographie, etc. La statistique spatiale commence à gagner du terrain dans de nouveaux domaines d'application qui nécessitent un traitement des données locales. En finance immobilière, l'intérêt porté à la localisation et à l'interaction spatiale commence à émerger vers le début des années 90 avec l'étude de Can (1990) qui note la présence d'une forte dépendance spatiale des prix immobiliers pouvant être traitée avec ces techniques. L'application de ces méthodes dans les études immobilières se développe peu à peu et en 1998, la revue de référence *The Journal of Real Estate Finance and Economics* consacre un volume spécial destiné à publier des articles spécifiques à ce thème.

Deux types de raisons peuvent expliquer ce développement de la statistique spatiale, du coté théorique et du coté empirique. Premièrement, le développement théorique rend possible l'ajustement fin de la dimension spatiale dans le modèle

d'estimation, permet l'exercice de la prévision et la formulation explicite de tests statistiques. Deuxièmement, du coté empirique, le développement des systèmes d'information géographique (SIG) permettent de collecter et de traiter les informations géographiques. Ces nouveaux outils ont fortement favorisé l'application des méthodes de la statistique spatiale.

Afin de mieux comprendre le principe de la statistique spatiale, elle peut être comparée avec l'étude des séries temporelles, plus utilisée et mieux connue dans le domaine de la finance. Si les techniques des séries temporelles permettent d'étudier la dynamique d'une variable dans temps, l'étude spatiale s'intéresse à la dynamique des réalisations dans l'espace. Avec la série temporelle, comme son nom indique, il s'agit de l'analyse d'une chronique historique : les variations d'une même variable au cours du temps. La réalisation de cette variable peut dépendre de sa réalisation passée ; ce type de dépendance temporelle des réalisations est dénoté par le terme d'autocorrélation temporelle. Les mêmes principes peuvent être appliqués avec les données réalisées dans l'espace. L'étude spatiale inclut dans la régression une variable indiquant la localisation de chaque réalisation. L'existence d'une dépendance entre des réalisations voisines est qualifiée d'autocorrélation spatiale. Le délai entre deux réalisations dans les séries temporelles est remplacé ici par la distance séparant chaque localisation. Si la technique des séries temporelles a pour principal objectif de déterminer les tendances, l'étude spatiale a pour objectifs de vérifier le degré de concentration, le taux de propagation des observations ainsi que la stabilité des valeurs dans l'espace.

Néanmoins, le problème de la dépendance spatiale paraît plus compliqué que celui de la dépendance temporelle. Trois principales complexités sont souvent mentionnées dans son étude. Premièrement, il n'y a pas d'ordre régulier dans l'espace comme dans le temps. L'analyse temporelle est appliquée avec un intervalle de temps qui est fixé et régulier, le modèle d'estimation peut être donc simplifié. Par contre en économétrie spatiale la distance entre chaque observation n'est pas nécessairement régulière, elle peut être variée de façon continue. Cependant, cette première difficulté peut être résolue en travaillant avec un intervalle de distance fixé en substitution dela distance réelle qui est une valeur continue. Deuxièmement, le temps ne possède qu'une seule dimension, par contre l'espace en a au moins deux. Ceci rend plus difficile le repérage de l'observation,

le modèle d'estimation et la prévision. De plus, l'étude spatiale peut aussi varier temporellement ; une analyse croisée entre la série temporelle et l'étude spatiale complique encore davantage les modèles d'estimation. Troisièmement, l'étude spatiale fait apparaitre non seulement la notion de distance, qui permet de définir le degré d'autocorrélation spatiale entre chaque observation, mais aussi la direction. Deux couples d'observations séparées par une même distance mais ayant des directions différentes peuvent avoir des degrés de corrélation spatiale qui ne sont pas identiques.

Les deux approches de la statistique spatiale sont l'économétrie spatiale et la géostatistique. Dans la littérature, ces deux approches sont utilisées pour étudier la dépendance spatiale des prix immobiliers. La géostatistique est appliquée aux données immobilières pour déterminer une estimation non biaisée du prix, pour donner une meilleure prédiction et pour déterminer les segmentations du marché immobilier. L'économétrie spatiale, qui semble mieux connue, est utilisée pour déterminer le degré de corrélation spatiale, estimer les indices ou pour évaluer la variation de prix liée à une externalité. Si ces deux méthodes sont adoptées pour étudier la dépendance spatiale des prix immobiliers dans plusieurs contextes variés, il n'existe cependant pas de règles très claires quant au choix de la méthode à employer. La question se pose donc : **Géostatistique – Économétrie spatiale : quelle approche pour quel contexte immobilier ?**

Cette question requière l'examen détaillé de ces deux approches : Quelles sont les distributions spatiales considérées ? Quelles sont les hypothèses posées ? Quelles sont les sources de l'autocorrélation spatiale considérée par chaque approche ? **Géostatistique – Économétrie spatiale : les ressemblances et les différences ? Les avantages et les inconvénients d'une approche par rapport à l'autre ?**

Les réponses à ces questions sont élaborées en cinq chapitres.

Le premier chapitre détaille les différents types de données spatiales et leurs caractéristiques spécifiques. Les deux problèmes principaux des données spatiales, l'autocorrélation spatiale et l'hétérogénéité spatiale, sont abordés. Plusieurs méthodes d'identification de l'autocorrélation spatiale sont présentées. Ce chapitre détaille aussi les différentes sources de l'autocorrélation, l'étude de la cause étant en effet une étape

préliminaire avant d'analyser la dépendance spatiale plus avant. Les deux approches de la statistique spatiale sont discutées brièvement dans ce chapitre. Pour terminer une revue de littérature financière est réalisée pour faire le point sur le développement et l'application de la statistique spatiale dans la recherche immobilière.

Le deuxième chapitre présente l'approche géostatistique, celle-ci estime directement la matrice de variance-covariance en supposant que la covariance entre les observations dépend inversement de la distance séparant leur localisation. La géostatistique est développée sur les hypothèses de continuité et de stationnarité du processus spatial. Cette approche s'applique à des données dont les localisations spatiales sont distribuées de façon aléatoire dans un espace continu. Ce chapitre est composé de deux parties. La première explique les fondements de l'approche géostatistique. Les hypothèses de continuité, de stationnarité spatiale ainsi que l'isotropie du variogramme sont énoncées. Le covariogramme et le semivariogramme, qui sont les deux fonctions centrales utilisées en géostatistique, sont présentés. Enfin, la méthode d'estimation paramétrique du variogramme ainsi que la méthode de prévision y sont aussi expliquées. Une deuxième partie étudie l'application de la géostatistique aux études immobilières. Plusieurs questions se posent à cette étape : Est-il raisonnable de considérer que les données immobilières sont distribuées continument dans l'espace sans tenir compte des frontières administratives ? Est-il raisonnable de supposer que la structure spatiale des données immobilières est stationnaire, que les prix des biens immobiliers ont le même comportement de dépendance spatiale s'ils se trouvent au centre ville ou en dehors de la ville ? Est-il raisonnable de supposer et d'imposer que la dépendance spatiale des valeurs immobilières dépende uniquement de la distance séparant leur localisation ? La direction a-t-elle de l'importance ?

Le troisième chapitre consiste en une étude empirique de la vraisemblance de l'hypothèse de stationnarité spatiale dans le cas des données immobilières. Ce chapitre estime le semivariogramme sur différents segments de marché et montre que la stationnarité est très discutable. Les semivariogrammes obtenus varient en effet significativement selon l'année d'étude et le segment de marché retenu. Il semble donc difficile de supposer que le processus associé à ces données soit globalement stationnaire. Ce problème de non-stationnarité nous conduit à tenter de déterminer les

paramètres qui pourraient remédier à cette variabilité dans l'espace. En ajoutant l'indicateur du quartier dans le modèle de régression hédonique, le semivariogram devient ainsi plus stationnaire. Ces résultats permettent de distinguer l'influence spatiale selon deux niveaux : l'effet de voisinage et l'effet de contiguïté

Le quatrième chapitre expose l'approche de l'économétrie spatiale, la deuxième approche de statistique spatiale. Elle définit et intègre la matrice d'interaction spatiale dans un modèle de régression. Cette matrice accorde des poids différents à chaque couple d'observations. Ces poids peuvent varier selon la distance entre les observations ou selon la contiguïté. La matrice de poids spatiaux est ensuite intégrée à l'équation de régression afin de prendre en compte la dépendance entre les observations voisines. La première partie de ce chapitre détaille plusieurs modèles d'économétrie spatiale existants et plusieurs façons de construire la matrice d'interaction spatiale. La deuxième partie du chapitre essaye de répondre à plusieurs questions : parmi les deux conditions utilisées pour déterminer la matrice d'interaction, quelle est la condition la plus appropriée à tel ou tel donnée ? Quelles sont les sources de l'autocorrélation spatiale adressées par chaque modèle ? Pourquoi seuls le modèle de dépendance spatiale des variables endogènes décalées et le modèle de l'autocorrélation spatiale des erreurs sont-ils choisis pour étudier les données immobilières dans la littérature ?

Le cinquième chapitre s'intéresse à une analyse de l'effet de diffusion, qui est une des sources de l'autocorrélation spatiale, et à la détermination du quartier dominant du marché immobilier d'une ville. Si la ville en question correspond à une structure mono-centrique, la réponse est sans doute le centre ville. En revanche, si cette ville correspond à une ville polycentrique, la réponse sera sans doute moins évidente. Comment définir alors économétriquement le quartier dominant ? Il n'est en général pas satisfaisant de l'identifier à partir d'un seul indicateur administratif. L'étude de l'effet de diffusion permet d'en discuter de la manière suivante : lorsque les observations d'un quartier sont enlevées de la base de données, si le niveau de corrélation spatiale estimé à partir des observations restantes baisse significativement comparé au niveau de corrélation spatiale estimé à partir de l'ensemble des données, ce quartier sera considéré comme un quartier dominant. Inversement, il est aussi possible que le niveau de corrélation spatiale augmente significativement lorsque les observations d'un quartier sont retirées de la base

de données ; ce quartier présente alors probablement des caractéristiques spécifiques qui ne se retrouvent pas dans les autres quartiers. Les résultats obtenus sont finalement comparés à la connaissance du terrain des professionnels de l'immobilier.

CHAPITRE I DONNEES IMMOBILIERES ET STATISTIQUES SPATIALES

1. Introduction

L'investissement immobilier prend une place de plus en plus prépondérante dans les choix de produits de placement des épargnants. Il permet de générer un revenu fixe régulier grâce au loyer et une plus-value provenant de la revente. L'immobilier est en ce sens semblable au produit obligataire. Malgré sa ressemblance avec le produit financier, l'immobilier présente certaines particularités. Notamment, les données immobilières possèdent de réelles caractéristiques spatiales: l'adresse, les coordonnées cartésiennes ou les coordonnées de latitude et longitude. En conséquence, le bien immobilier ne peut pas être évalué comme un simple produit financier. Son prix ne dépend pas uniquement de ses caractéristiques physiques, mais aussi de ses caractéristiques de localisation. Par ailleurs, à l'instar de toutes les données financières, les prix immobiliers sont collectés au fils du temps et peuvent parfois présenter une dépendance temporelle. Mais, étant donné qu'ils présentent en outre ces caractéristiques géographiques, les prix immobiliers se différencient de prix des autres produits financiers par le fait qu'ils exposent en plus au problème de dépendance spatiale.

Can (1990) est l'un des premiers articles à mentionner le problème de corrélation spatiale des prix entre biens voisins. Plusieurs raisons sont fournies pour expliquer cette dépendance. D'une part, les biens voisins sont similaires parce qu'ils sont construits sur la même période et se trouvent au sein du même environnement. D'autre part, le processus de valorisation du bien immobilier implique lui-même dès le départ cette dépendance : le propriétaire se renseignant bien souvent auprès de ses voisins pour fixer la valeur de son bien. En outre, certaines variables explicatives omises dans le modèle d'évaluation immobilière peuvent aussi causer une dépendance spatiale. En cas de présence de corrélation spatiale, le modèle hédonique classique utilisant les moindres carrées ordinaires fournit des estimations biaisées et non efficientes. En effet, cette particularité des données immobilières nécessite de prendre en compte la dépendance spatiale dans le modèle d'évaluation des biens immobiliers. Certaines études passées ont cherché à éliminer la dépendance spatiale en incluant les caractéristiques spatiales dans le modèle d'estimation des prix hédoniques. Pourtant, malgré le nombre important de variables explicatives ajoutées au modèle, les problèmes d'autocorrélation spatiale et d'instabilité des paramètres estimés selon la localisation persistent encore. Afin de mieux intégrer le

problème de dépendance spatiale, l'outil statistique utilisé doit se révéler adapté à la nature spatiale des variables traitées. La statistique spatiale apparaît donc comme une méthode adaptée pour analyser les données distribuées dans l'espace qui présentent le problème de corrélation spatiale.

La statistique spatiale est initialement utilisée dans les domaines des sciences de l'environnement, sciences de la terre, agronomie ou météorologie. Elle commence à gagner du terrain dans l'étude économique et financière des années 90s. Récemment, grâce au progrès des technologies de positionnement (le Système d'Information Géographique (SIG)) qui permet de faciliter la collecte des données, et au développement théorique de la méthode qui permet de faciliter son application dans des domaines plus variés, la statistique spatiale prend une place de plus en plus importante dans l'étude immobilière.

Selon la littérature, les deux approches de statistique spatiale souvent utilisées pour étudier la dépendance spatiale des prix immobiliers sont la géostatistique et l'économétrie spatiale. Ces deux approches sont distinctes : chacune s'applique à des distributions de données différentes et se base sur des hypothèses contraignantes distinctes. La géostatistique est développée afin de s'adapter à la distribution des observations dans un espace continu, comme l'évolution de gisement minier, son premier champ d'application. L'hypothèse de la stationnarité spatiale est supposée, afin de permettre l'analyse globale des observations. A l'inverse, l'économétrie spatiale est développée sous l'hypothèse d'une répartition des données de type treillis ou latticiel. Cette approche nécessite un choix *ex ante* du modèle d'estimation en fonction de la source de la dépendance spatiale.

Ce chapitre est destiné à donner une brève synthèse sur les données spatiales et leurs particularités, à détailler le problème de l'autocorrélation spatiale souvent rencontré en étude immobilière et à présenter les efforts déployés pour résoudre ce problème. Les différents types de données spatiales sont présentés dans la deuxième section. Etant donné que les deux approches de statistique spatiale sont basées sur des hypothèses de distribution des données différentes, il est nécessaire de connaître les différents types des données spatiales. Le choix de la distribution des données oriente le choix de l'approche à utiliser. La distance entre les couples d'observations est un élément important pour une

étude spatiale, les méthodes de calcul de la distance sont donc présentées dans la même section. Une fois les données spatiales collectées, le problème le plus souvent rencontré est la dépendance spatiale. Une étape préliminaire avant d'appliquer l'étude statistique est donc de vérifier la présence de cette dépendance. Les tests de présence d'autocorrélation spatiale sont donc présentés dans la troisième section. La quatrième section est destinée à montrer l'application de la statistique spatiale à une estimation hédonique. Une courte présentation de l'estimation hédonique ainsi que des différentes sources de dépendance spatiale des prix hédoniques sont présentées dans cette section. Dans la cinquième section, les méthodes utilisées pour analyser les données spatiales sont détaillées. L'application de la statistique spatiale peut être faite de deux façons : premièrement, étudier la covariance des résidus en passant par le modèle *géostatistique* ou, deuxièmement, réévaluer le modèle en incorporant la matrice des points spatiaux qui permet de prendre en compte la dépendance entre les observations en passant par le modèle d'économétrie spatiale. Ce chapitre ne donne qu'une courte présentation des deux approches. Les différentes hypothèses contraignantes, ainsi que l'applicabilité à l'immobilier sont présentées dans les chapitres suivants. Ce chapitre se termine par la sixième section concernant la revue de la littérature sur l'application de la statistique spatiale à l'étude immobilière.

2. Données spatiales et leurs particularités

Afin d'étudier les données spatiales, il est nécessaire de connaître la distribution des observations dans l'espace. Cette distribution spatiale peut être désignée grâce à l'information sur les références spatiales. De plus, cette information de localisation permet de calculer la distance entre chaque observation, ce qui est un élément indispensable pour déterminer l'ensemble des observations voisines ayant servi à analyser la dépendance spatiale. Les observations spatiales sont confrontées à deux spécificités : *l'autocorrélation spatiale* qui se réfère à l'absence d'indépendance entre observations géographiques et *l'hétérogénéité spatiale* qui indique la différenciation dans l'espace des variables et des comportements. La suite de cette étude se concentre essentiellement sur le problème de l'autocorrélation spatiale.

2.1. Données spatiales

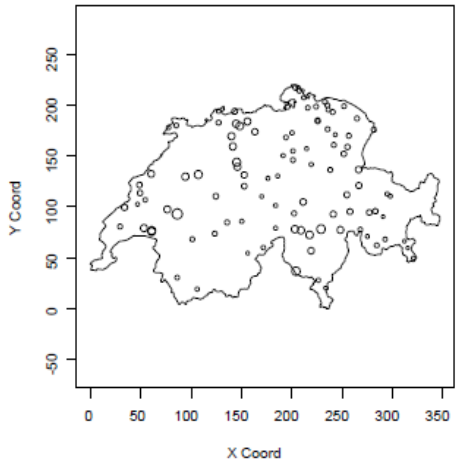
Les données spatiales sont constituées d'observations auxquelles est associée une information géographique indiquant sa localisation. Jayet (1993) précise que la référence spatiale d'une donnée statistique peut être révélée sous deux formes principales. D'une part, les informations géographiques ponctuelles indiquent les points particuliers répartis dans l'espace. Dans le cas des données de l'étude immobilière, ce sont les informations du type : position sur une carte, position dans un référentiel géographique, coordonnées cartésiennes, etc. D'autre part, ces informations peuvent être agrégées. Les taux de criminalité d'un arrondissement ou le niveau de pollution d'un quartier peuvent être par exemple considérés comme des informations spatiales. Cette section s'intéresse seulement aux indicateurs ponctuels et montre les problèmes souvent rencontrés sur les données spatiales et comment les analyser.

Soit $s \in \mathbb{R}^d$, les localisations dans l'espace euclidien de dimension d. Si $d = 2$, dans ce cas, l'indice de localisation $s = (x, y) \in \mathrm{D}$ est repéré par le coordonnée géographique (latitude-longitude ou coordonnée cartésienne). $Z(s)$ indique les observations collectées à la localisation s avec s variant dans le sous-espace $D \subset \mathbb{R}^d$. La réalisation d'une variable aléatoire indexée par la localisation spatiale s et appartenant à un ensemble spatial D est définie comme : $z = \{z(s), s \in D\}$. Les différents types de données statistiques sont définis par rapport à la répartition de s dans D. La localisation d'un site d'observation $s \in D$ est soit fixée et déterministe, soit aléatoire, ce qui donne alors deux types de données spatiales et deux approches correspondant à chaque type des données.

2.1.1. Données géostatistiques

Les données sont de type géostatistique si D est un sous espace *continu* de $\mathbb{R}^d$. L'indice de localisation s varie de façon *aléatoire* dans le sous ensemble de $\mathbb{R}^d$. La réalisation $z(s) \in \mathbb{R}$ est la valeur réelle observée en n sites fixés $\{s_1, \dots, s_n\} \subset D$ (Gaetan, Guyon et Bleakley (2010)).

Figure I.1 : L'exemple de la distribution aléatoire irrégulière des données de type géostatistique



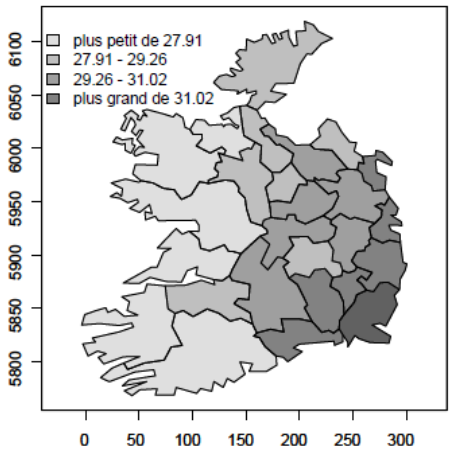
Cumuls pluviométriques sur le réseau météorologique suisse le 8 mai 1986
(Source : Gaetan, Guyon and Bleakley (2010))

La Figure I.1 illustre bien la distribution des données dans l'espace. Cette distribution aléatoire (irrégulière) concerne des données géostatistiques. Droesbeke, Lejeune et Saporta (2006a) indiquent que ces données sont qualifiées de « variables régionalisées » par Matheron (1963), le fondateur de la géostatistique. Ce dernier propose une approche géostatistique utilisée pour analyser la dépendance spatiale avec deux outils de base : la fonction de covariance (le *covariogramme*) et le *semivariogramme*. Le gisement minier, la pollution atmosphérique ou la distribution des pluies sont des exemples de données géostatistique dont la distribution des données dans l'espace continu est supposée aléatoire.

2.1.2. Données latticielles

Dans le cas des données de type treillis, aussi appelées données latticielles (LeSage et Pace (2009)), *D* est un sous espace *discret fixé* et contient une collection dénombrable de points de $\mathbb{R}^d$. L'indice de la localisation *s* est donc *fixé et déterministe. Il* représente par exemple des unités géographiques d'un réseau structuré ou des segmentations administratives.

Figure I.2 : L'exemple de la distribution régulière des données de type treillis



Pourcentage d'individus du groupe sanguin A dans les 26 comtés de l'Irlande
(Source : Gaetan, Guyon and Bleakley (2010))

La Figure I.2 présente la distribution dans l'espace des données de type treillis. La comparaison entre la Figure I.1 et la Figure I.2 illustre bien la différence de distribution des données. Cressie (1991) explique que l'information satellite de la distribution minérale ou de la qualité de l'air peuvent aussi être considérés comme les données latticielles, selon ce type d'information la terre est divisée en plusieurs petites parties (pixels) et le point représentant chaque information se trouve au centre de chaque pixel. Par conséquent, les informations obtenues sont régulières dans la surface de $\mathbb{R}^2$. Les données latticielles peuvent être comparées aux données d'une série temporelle avec un écart du temps constant. L'économétrie spatiale est l'approche souvent appliquée pour analyser ce type de données. Cette approche estime la dépendance spatiale à l'aide de la matrice de poids spatiaux prenant en compte la relation entre des observations voisines.

2.1.3. Données ponctuelles (point patterns)

Le troisième type de données qui est souvent mentionné dans la littérature, mais ne sera pas traité dans cette étude est les données ponctuelles. Les observations sont la réalisation aléatoire d'un processus ponctuel et ces réalisations sont observées en positions spatiales aléatoires. L'étude des données ponctuelles est par exemple la

répartition de la détection des cas de grippe aviaire chaque mois, cette répartition des observations étant aléatoire et variant mensuellement. L'étude des données ponctuelles repose souvent sur l'étude du comportement de la réalisation (réalisation régulière, groupée ou concentrée autour d'un certain point de repère).

La distinction entre ces différents types de données est parfois difficile. Les données latticielles dont la taille de pixel est très petite peuvent être considérées comme ayant une répartition continue donc comme des données géographiques. Les données géostatistiques peuvent être transformées en données latticielles si le sous espace est découpé en différents pixels et si le composant de chaque pixel est défini comme la réalisation observée. Cependant, connaître le type de données permet de mieux sélectionner la méthode la plus appropriée ainsi que fixer un objectif correspondant au mieux aux données étudiées. Le type de données a aussi une importance dans la détermination des voisinages. La contiguïté (existence d'une frontière en commun ou non) est celle la plus utilisée pour travailler avec les données latticielles. Par contre, il paraît difficile de définir la frontière commune pour le voisinage, si les données sont de type géostatistique et sont réparties de façon continue dans l'espace.

2.2. Localisation et calcul de la distance

Afin de pouvoir appliquer une étude spatiale, les deux éléments nécessaires sont l'indicateur de localisation de chaque observation et la distance entre chaque couple d'observations. L'indicateur de localisation est la référence dans l'espace, et peut être présenté sous la forme de coordonnées géographiques ou de coordonnées cartésiennes. Cet indicateur est nécessaire pour calculer, par la suite, la distance entre deux points dans l'espace. Plusieurs méthodes de calcul de la distance sont disponibles : la distance euclidienne, la distance de placement ou la distance du grand cercle. Chaque méthode donne un résultat sensiblement différent, le choix de la méthode dépendant directement de l'objectif de l'étude, ainsi que des informations disponibles dans la base de données.

2.2.1. Localisation

Les données spatiales doivent être référencées par une localisation précise afin de pouvoir calculer la distance entre chaque observation. Les indicateurs de localisation peuvent être les coordonnées géographiques, les coordonnées cartésiennes ou l'adresse exacte de chaque observation dans le cas de l'étude immobilière. Les coordonnées géographiques (les couples latitude et longitude) ou les coordonnées cartésiennes (les couples latitude et longitude projetées) sont les informations les plus convenables pour calculer la distance. La base de données spatiales dispose normalement d'un de ces deux indicateurs car la conversion entre ces deux indicateurs est possible. A partir de la latitude et la longitude, les coordonnées cartésiennes sont obtenues en projetant les coordonnées sphériques dans un espace cartésien à deux dimensions. Plusieurs techniques de projection cartographique ont été développées par le modèle mathématique : la projection cylindrique, la projection conique et la projection azimutale.

Un des problèmes majeurs de l'étude spatiale en immobilier est que certaines bases de données ne disposent d'information ni sur des coordonnées géographiques ni sur des coordonnées cartésiennes. La localisation d'un bien est uniquement indiquée par l'adresse du bien. Les logiciels de géo-localisation comme *Yahoo map*[1], *Geocoder* ou *Google earth* sont les moyens permettant d'obtenir les coordonnées géographiques d'un bien grâce à son adresse.

2.2.2. Calcul de la distance

La distance est une deuxième information nécessaire pour une étude de statistique spatiale, que ce soit *via* le modèle de l'économétrie spatiale ou le modèle géostatistique. La distance entre deux observations est utilisée comme une pondération indiquant l'importance qu'une observation a sur une autre. Cette pondération est normalement l'inverse de la distance ou l'inverse de la distance au carré. Plus l'observation est éloignée, moins l'on accorde de poids à cette observation. Dans le cas de la recherche en immobilier, la distance utilisée peut être la distance par rapport au quartier des affaires, la

[1] Cette étude développe le code VBA utilisé pour obtenir les coordonnées géographiques à partir de l'adresse en passant par le logiciel *Yahoo map*.

distance jusqu'au transport en commun ou la distance jusqu'à l'école, etc. Ces informations sont souvent ajoutées dans la régression pour permettre de mesurer l'effet de la localisation sur le prix immobilier.

Dans l'espace à deux dimensions, le théorème de Pythagore est la méthode la plus utilisée pour calculer la distance entre deux points, du fait de sa facilité d'élaboration. Mais elle présente certaines contraintes d'utilisation. La distance sphérique est proposée dans le cas où la base de données dispose des longitudes et des latitudes. Cette distance sphérique est normalement utilisée dans le cas où les données sont séparées par une distance importante dont le calcul nécessite de prendre en compte la courbure de la terre. La méthode la plus appropriée dépend de l'information de la localisation disponible dans la base de données et de l'objectif de l'analyse.

Distance en deux dimensions

La **distance euclidienne** est la ligne droite, la plus courte, qui relie deux points dans l'espace à deux dimensions. C'est la méthode la plus simple pour calculer la distance entre deux points, mais il est nécessaire d'avoir les coordonnées cartésiennes de chaque observation. Cette méthode est utilisée pour calculer une petite distance, par exemple le cas d'une étude de botanique dans un terrain limité. En immobilier, cette méthode peut être utilisée pour calculer la distance entre deux observations localisées dans le même quartier ou la même ville. Le théorème de Pythagore est appliqué pour retrouver la distance euclidienne, la distance *d* étant définie par :

$$d_e = \left[(x_i - x_j)^2 + (y_i - y_j)^2\right]^{1/2} \qquad \text{Eq. I.1}$$

avec (x_i, y_i) et (x_j, y_j) indiquant la localisation cartésienne des deux points (S_i et S_j) dans l'espace à deux dimensions, et pouvant s'exprimer en n'importe quelle unité (miles, kilomètres, yards…). L'indice *e* signifie la distance euclidienne.

Dans le cas d'une étude d'urbanisme, si l'objectif de l'étude est de calculer la **distance de déplacement** des habitants, la distance la plus appropriée n'est pas la distance euclidienne. Dans cette situation, c'est plutôt la distance de transport ou la

distance du chemin de placement qui convient le mieux à l'objectif de l'analyse. Sous l'hypothèse que plusieurs grandes villes soient construites en forme de bloc et que les routes suivent un système de grille, la distance de placement (d_p) entre deux points peut donc être définie par :

$$d_p = |x_i - x_j| + |y_i - y_j| \qquad \text{Eq. I.2}$$

avec $|x_i - x_j|$ désignant la longueur en valeur absolue entre les points x_i et x_j, et $|y_i - y_j|$ désignant la longueur en valeur absolue entre les points y_i et y_j. Au lieu de prendre en compte la longueur de l'hypoténuse, cette méthode considère la somme des longueurs des deux autres côtés.

Les calculs de la distance euclidienne et de la distance de placement peuvent être généralisées par **la distance Minkowski** (d_M) (Kruskal (1964)) :

$$d_M = \left(|x_i - x_j|^k + |y_i - y_j|^k\right)^{1/k} \qquad \text{Eq. I.3}$$

Selon l'Eq. I.3, le cas où $k = 1$, nous retrouvons la distance de placement (Eq. I.2) et si $k = 2$ cette équation donne la distance euclidienne (Eq. I.1). La valeur de *k* peut aussi être comprise entre 1 et 2. Supposons que la distance réelle de placement soit supérieure à la droite directe (l'hypoténuse du triangle), mais plus courte que la somme des deux côtés du triangle, la valeur de *k* se trouvera entre 1 et 2. Remarquons que la régression entre la distance observée et la distance calculée peut être utilisée pour déterminer la valeur de *k* la plus appropriée.

Distance du grand cercle

La distance euclidienne est normalement utilisée dans le cas où les coordonnées cartésiennes sont disponibles. Par contre, certaines bases de données, surtout en immobilier, ne disposent que des latitudes et longitudes comme indicateur de localisation. De plus, la distance euclidienne est notamment utilisée pour calculer la distance dans un petit espace. Dans le cas d'une grande distance -par exemple la distance entre deux villes dans deux continents- la distance du grand cercle -appelée distance orthodromique- est

plus appropriée car permet de tenir compte de la courbure de la terre. Cette distance du grand cercle correspond à un arc reliant deux points sur une sphère. Par conséquent, la distance du grand cercle est calculée en prenant en compte les coordonnées de latitude et longitude (en radian) et le rayon de la terre.

Il existe deux façons de calculer la distance du grand cercle : **la loi du cosinus** (d_C) et **la formule *Haversine*** (d_H).

$$d_C = R \times \arccos\left[\sin x_i \cdot \sin x_j + \cos x_i \cdot \cos x_j \cdot \cos(y_j - y_i)\right] \qquad \text{Eq. I.4}$$

et

$$d_H = 2R \times arcsin\sqrt{\sin^2\left(\frac{x_i - x_j}{2}\right) + + \cos x_i \cdot \cos x_j \cdot \sin^2\left(\frac{y_j - y_i}{2}\right)} \qquad \text{Eq. I.5}$$

avec $(x_i, y_i), (x_j, y_j)$ indiquant la latitude et la longitude d'une bien localisé au point S_i et S_j et R égal à 6371 kilomètres ou 3959 miles correspondant au rayon de la terre.

Les latitudes et les longitudes sont souvent exprimées en système sexagésimal (degré, minute et seconde). Par ailleurs pour appliquer la formule de la distance du grand cercle, il faut d'abord transformer la coordonnée du système sexagésimal en système décimal puis la diviser par $\pi/180$, pour finalement obtenir la coordonnée exprimée en radian. Ces deux formulations sont mathématiquement équivalentes, mais la formule *Haversine* est plus utilisée parce que la loi du cosinus peut générer des erreurs d'arrondis si elle est utilisée pour calculer la distance entre deux points très proches. Néanmoins, la formule *Haversine* peut aussi causer une erreur dans le cas de calcul de la distance entre deux points opposés de la terre.

Il faut également noter que la formule de la distance du grand cercle est basée sur certaines hypothèses. Premièrement, le calcul suppose que la terre est de forme sphérique. En réalité, la terre a la forme d'un ellipsoïde, légèrement aplati aux pôles. Deuxièmement, la topographie du terrain -comme la présence d'une rivière ou de la montagne- n'est pas prise en compte dans cette formule, alors qu'elle peut représenter un facteur important dans le calcul de certaines distances.

Cette section a présenté seulement quelques formules utilisées pour calculer la distance entre deux localisations. Il existe cependant d'autres méthodes qui paraissent plus pertinentes, mais ne sont pas présentées ici à cause de leur limite d'application. La distance routière entre deux localisations donne par exemple une information précise mais demande une cartographie très détaillée. La distance peut aussi être exprimée comme le temps nécessaire à parcourir un trajet d'un point à l'autre. Déterminer le temps nécessaire au placement demande toutefois un logiciel bien spécifique.

Plusieurs logiciels proposent le calcule automatique des différentes mesures de la distance comme le Système d'information géographique (SIG) ou *Google Earth*. Cela donne une facilité de calcul, mais le choix de la distance à calculer est toutefois nécessaire pour définir la mesure la plus appropriée à l'objectif de l'analyse et la plus adaptée à la base de données existante.

2.3. Autocorrélation spatiale et hétérogénéité spatiale

L'étude statistique suppose souvent que les observations sont des variables indépendantes et identiquement distribuées (*iid*), ce qui conduit à supposer que les données sont des variables aléatoires, ont toutes la même loi de probabilité et sont mutuellement indépendantes. Si la base de données est non spatiale, la relation linéaire entre les n observations indépendantes $y_i, i = 1, \dots, n$, et les variables explicatives x_{ik} regroupées dans la matrice X peut être représentée par l'équation suivante :

$$y_i = X_i\beta + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2) \qquad \text{Eq. I.6}$$

X_i représente le vecteur des variables explicatives de dimension $1 \times k$ et β le vecteur des coefficients estimés lors de la régression de dimension $k \times 1$. Selon l'Eq. I.6, les résidus de l'estimation ε_i , $i = 1, \dots, n$ sont indépendants et suivent identiquement la loi normale $N(0, \sigma^2)$ de moyenne nulle et d'écart-type constant σ^2 . L'hypothèse d'indépendance des observations se traduit par l'absence de relation entre les résidus de l'estimation. Statistiquement, cela se représente par $Cov(\varepsilon_i \varepsilon_j) = 0$ pour $i \neq j$ donc $E(\varepsilon_i \varepsilon_j) = E(\varepsilon_i)E(\varepsilon_j) = 0$.

Dans le cas des données spatiales, l'indicateur *i* représente, cette fois-ci, la région ou la localisation d'où les observations sont collectées. Les données spatiales sont indépendantes et non corrélées si la valeur collectée à l'endroit S_i ne dépend pas de celle collectée à l'endroit S_j . Pour l'étude de l'immobilier physique, cette hypothèse de simplification n'est pas toujours respectée. Le prix immobilier dépend fortement de sa localisation et de son environnement. Cela conduit notamment au problème d'autocorrélation spatiale des prix de biens voisins. L'effet de voisinage et l'effet de contiguïté ont une influence importante dans l'estimation des valeurs immobilières. Les analyses précédentes montrent que les données spatiales convergent souvent sur deux aspects spatiaux : *l'autocorrélation spatiale* définie par la présence de dépendance spatiale des observations voisines et *l'hétérogénéité spatiale* qui consiste dans la différenciation des variables et des comportements dans l'espace.

2.3.1. Autocorrélation spatiale

Selon la littérature, l'autocorrélation spatiale peut être définie de plusieurs façons.

Anselin (1988) définit l'autocorrélation spatiale comme la coïncidence entre valeur similaire et localisation similaire ([…] *the coincidence of value similarity with locational similarity*).

LeSage et Pace (2009) indiquent que l'autocorrélation apparaît quand les valeurs observées à une localisation précise dépendent des observations voisines ([…] *a situation where values observed at one location or region depend on the values of neighboring observations at nearby locations*).

Droesbeke, Lejeune et Saporta (2006b) expliquent que la statistique spatiale se différencie de la statistique classique par des observations analysées généralement dépendantes, cette dépendance étant due à la localisation relative des points de mesure.

En effet, l'autocorrélation spatiale se réfère à la ressemblance des observations en fonction de leur localisation géographique. La dépendance spatiale ou l'autocorrélation spatiale se présente si l'observation d'une variable est spatialement corrélée avec celles reparties dans l'espace en des localisations différentes. Plus précisément, cette

dépendance spatiale se trouve, normalement, parmi les observations voisines et l'absence de l'autocorrélation signifie qu'il n'existe pas de ressemblance entre les observations voisines. Supposons que la valeur observée à un endroit S_i (ou une région), notée observation y_i, soit spatialement corrélée à sa valeur voisine collectée à l'endroit S_j, notée observation y_j. Formellement, la valeur y_i peut être définie par :

$$y_i = f(y_j) \qquad \text{Eq. I.7}$$

Plus précisément,

$$\begin{aligned} y_i &= \alpha y_j + X_i\beta + \varepsilon_i \\ y_j &= \alpha y_i + X_j\beta + \varepsilon_j \\ \varepsilon_i, \varepsilon_j &\sim N(0, \sigma^2) \end{aligned} \qquad \text{Eq. I.8}$$

L'Eq. I.8 montre que la réalisation de l'observation y_i dépend de celle de y_j parce que les localisations S_i et S_j sont les points voisins dont la covariance vérifie $Cov(y_i, y_j) \neq 0$ pour $i \neq j$. Confrontée au problème de dépendance spatiale des variables, la méthode statistique utilisée doit se révéler adaptée à la nature spatiale des variables traitées. La non prise en compte de la dépendance spatiale dans la modélisation peut occasionner des problèmes de mauvaise spécification et de biais d'estimation (Anselin (1988)).

Dans le cas de l'immobilier physique, l'autocorrélation spatiale des prix immobiliers est souvent citée. Plusieurs raisons peuvent être invoquées pour expliquer cet aspect spatial. Tout d'abord, les valeurs immobilières dépendent partiellement de leur localisation. Comme le voisinage a souvent la même structure, la même qualité et la même période de construction, cela peut créer un problème d'autocorrélation des prix. Le processus d'évaluation immobilière, soit par le propriétaire, soit par l'expert immobilier, prend comme référence les valeurs voisines. Il est donc possible de trouver de la dépendance parmi les prix voisins. L'explication plus détaillée sur les sources de l'autocorrélation spatiale sera présentée dans le point 4.2 de ce chapitre. Généralement, l'autocorrélation peut être négative ou positive, mais il paraît plus raisonnable qu'elle soit positive dans le cas de l'étude immobilière, c'est-à-dire les observations collectées des localisations voisines se ressemblent davantage qu'elles ne ressemblent aux autres valeurs. L'existence d'un parc naturel crée une autocorrélation positive des prix des biens

immobiliers localisés proche de ce parc. Les valeurs de biens immobiliers localisés proche du périphérique sont positivement corrélées parce que ces valeurs sont affectées négativement par la présence de pollution et de nuisances sonores.

L'autocorrélation spatiale peut être comparée à l'autocorrélation temporelle. Néanmoins, le problème de la dépendance spatiale paraît plus compliqué que celui de la dépendance temporelle. Trois sources principales de complexité sont souvent mentionnées dans l'étude de la dimension spatiale. Premièrement, il n'y a pas d'ordre régulier dans l'espace, à la différence du temps. L'analyse temporelle est appliquée avec un intervalle de temps qui est fixé et régulier, le modèle d'estimation peut donc être simplifié. Par contre, en économétrie spatiale, la distance entre chaque observation n'est pas nécessairement régulière. De plus, elle peut varier de façon continue. Cependant, cette première difficulté peut être résolue en travaillant avec un intervalle de distance fixé à la place de la distance réelle qui est une valeur continue. Deuxièmement, l'autocorrélation temporelle est une fonction unidirectionnelle puisque seul le passé influence le futur. Par contre, l'autocorrélation spatiale est un cas multidirectionnel : tout est relié à tout. Cela rend plus difficile d'élaborer l'indicateur de l'observation, le modèle d'estimation et la prévision. En outre, l'étude spatiale peut aussi varier temporellement, l'analyse croisée entre la série temporelle et l'économétrie spatiale compliquant d'autant plus le modèle d'estimation. Troisièmement, l'étude avec la dimension spatiale fait apparaitre non seulement la notion de distance, qui permet de définir le degré d'autocorrélation spatiale entre chaque observation, mais aussi la direction. Deux couples d'observations séparées par la même distance, mais en différentes directions, peuvent avoir un degré de corrélation spatiale différente. Ces trois complexités peuvent donc expliquer l'utilisation moins connue de l'économétrie spatiale.

Reprenons le système d'équation Eq. I.8 et supposons que toutes les observations soient reliées. S'il y a n observations collectées à n localisations, il existe $n^2 - n$ relations de dépendance entre les variables, donc au moins $n^2 - n$ paramètres à estimer. Ces paramètres à estimer peuvent être réduits en imposant une structure d'autocorrélation spatiale entre les observateurs. Par exemple, les observations sont dépendantes uniquement si la distance entre elles ne dépasse pas une certaine distance limite ou si les observations se localisent dans le même sous espace ou le même quartier. Cette structure

de dépendance spatiale peut être déterminée par la portée de covariogramme, c'est-à-dire la distance à partir de laquelle les observations peuvent être considérées comme indépendantes de la localisation géographique, pour l'approche géostatistique, ou par la condition de voisinage pour l'approche de l'économétrie spatiale.

Afin de traiter le problème de l'autocorrélation spatiale, deux approches méthodologiques sont souvent utilisées : soit l'économétrie spatiale avec la matrice de pondérations spatiales, soit la géostatistique avec le covariogramme ou semivariogramme. Les éléments distinctifs, les avantages et inconvénients ainsi que le champ d'application possible de chaque approche sont commentées de façon détaillée dans le CHAPITRE II et le CHAPITRE IV.

2.3.2. Hétérogénéité spatiale

La deuxième particularité des données spatiales est l'hétérogénéité spatiale. Ce phénomène consiste en une instabilité des relations entre des observations dans l'espace. Cela conduit donc à l'instabilité des paramètres de la régression. Gallo (2002) indique qu'en pratique, ces différences peuvent se traduire de deux façons dans une régression: par des coefficients différents ou par des variances des termes d'erreurs différentes selon la localisation. Dans le premier cas, cela correspond à l'instabilité des paramètres de la régression, ces paramètres estimés pouvant varier selon la localisation étudiée. Le second cas est le problème de l'hétéroscédasticité. En cas d'absence de stabilité des relations entre les variables, le modèle de régression doit permettre de prendre en compte les caractères particuliers de chaque localisation étudiée, soit en laissant varier les paramètres estimés selon la localisation, soit en supposant une relation différente pour chaque zone géographique.

Formellement, l'équation de régression dans le cas de présence de l'hétérogénéité spatiale est la suivante :

$$y_i = X_i\beta_i + \varepsilon_i \qquad \text{Eq. I.9}$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

avec i les indicateurs des observations collectées au point dans l'espace $i = 1, \dots, n,\ X_i$ vecteur des variables explicatives de dimension $(1 \times k)$, β_i le coefficient de la régression, y_i les variables endogènes indépendantes et ε_i indiquant les résidus de l'estimation. L'Eq. I.9 diffère de l'Eq. I.6 par le coefficient d'estimation β_i. L'indice i pour la valeur de β signifie que le vecteur des paramètres est différent pout chaque observation i. La valeur β_i varie d'une localisation à une autre.

Dans le cas de l'étude immobilière, il paraît possible que l'impact spatial ne soit pas homogène. Can (1990) pose la question : « la valeur des caractéristiques des biens immobiliers varie-t-elle selon la localisation du bien ? ». Si cette valeur varie, une seule estimation hédonique ne suffit pas pour estimer l'ensemble des observations. Il faut un ou des indicateurs qui permettent de faire varier l'estimation hédonique selon la localisation des biens. Plusieurs exemples peuvent être évoqués pour confirmer sa remarque. La caractéristique de localisation « vue sur Sienne » est une variable qui a une influence possible sur le prix uniquement pour les appartements situés au bord de la Sienne. En revanche, cette caractéristique n'a aucune influence sur le prix des appartements situés à coté de l'avenue des Champs-Elysées. Inversement, la caractéristique « vue sur Champs-Elysées » influence le prix uniquement pour les appartements situés à coté de cette avenue et non pas pour ceux situés au bord de la Sienne. Les caractéristiques de localisation des biens immobiliers diffèrent selon leurs environnements. Un seul modèle d'évaluation avec les mêmes caractéristiques pour toutes les segmentations donne résultat biaisé. Par conséquent, les caractéristiques de localisation prises en compte dans le modèle d'évaluation doivent varier selon la segmentation de marché. Prenons l'exemple de la distance jusqu'au transport en commun. Etre situé près d'un accès à un transport en commun augmente la valeur d'un appartement grâce à la facilité de déplacement, mais peut avoir une influence moins importante voire même négative sur la valeur d'une maison. Les acheteurs de maison disposant souvent de voiture, la localisation près d'un accès de transport en commun peut être perçue, pour eux, comme un inconvénient du fait de la nuisance sonore ou de la fréquentation des passagers. Un autre exemple peut-être donné sur l'existence de jardin. Cette variable donne une influence différente selon la localisation des maisons. Si la maison se trouve au centre ville, l'existence de jardin augmente significativement la valeur du bien. En revanche si la maison se trouve en milieu rural, il paraît normal qu'une maison dispose un jardin, donc

que cette caractéristique soit moins valorisée. Dans ce cas, le paramètre de la régression lié à l'existence de jardin varie selon la localisation.

Même si le problème d'hétérogénéité spatiale paraît important pour une analyse des données spatiales, il est beaucoup moins mentionné que l'autocorrélation spatiale par la littérature dans l'étude immobilière. Par ailleurs, comme l'objectif de cette étude est le traitement du problème de l'autocorrélation spatiale, l'hétérogénéité spatiale est donc évoquée brièvement.

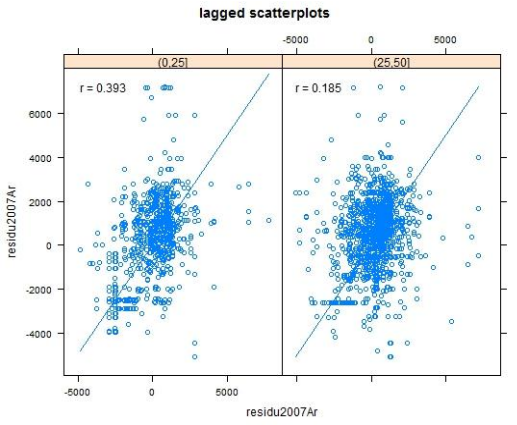
3. Test de l'autocorrélation spatiale

Comme précédemment indiqué par Gallo (2002), les données spatiales sont normalement supposées indépendantes alors que cette hypothèse est rarement justifiée et devrait être systématiquement testée. Une fois collectée les données spatiales, la condition d'indépendance spatiale doit donc être, tout d'abord, vérifiée. Plusieurs méthodes ont été proposées pour vérifier cette condition ; le *H-Scatterplot*, le covariogramme et l'indice de Moran.

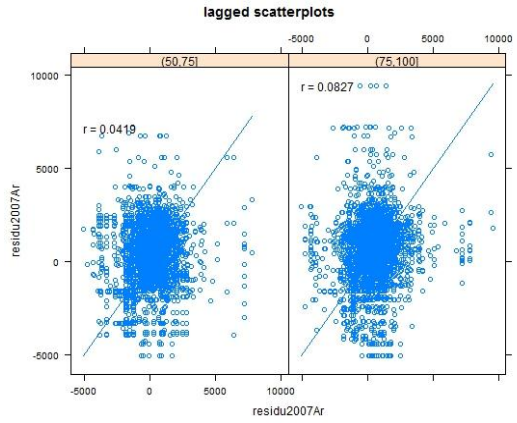
3.1. H-Scatterplot

Un *H-Scatterplot* permet de montrer toutes les paires possibles des observations dont les emplacements sont séparés par une certaine distance *h*. Présenté dans un plan *(x, y)*, l'axe des abscisses est libellé y_i référée à l'observation à la position S_i et l'axe des ordonnées est libellé y_j référée à l'observation à la position $s_j = s_i + h$ pour chaque distance *h* considérée. La forme du nuage de points montre la continuité des données sur une certaine distance dans une direction particulière. Si les données dont la position est séparée par la distance *h* sont très semblables (très corrélées), le nuage de points se trouve autour de la ligne $x = y$, une droite à 45 degrés passant par l'origine. Dans le cas où les données sont moins corrélées, le nuage de points sur le *h-Scatterplot* devient plus dispersé et plus diffus.

Figure I.3 : Les 4 *H-Scatterplots* des résidus obtenus par la régression hédonique



(A) (B)



(C) (D)

Les résidus obtenus de la régression linéaire entre les valeurs de transactions des appartements parisiens en 2007 et leurs caractéristiques physiques. Les graphiques (A) (B) (C) et (D) représente les différentes intervalles utilisées : (0 ; 25], (25 ; 50], (50 ; 75] et (75 ; 100].

La Figure I.3 présente les 4 *H-Scatterplots* des résidus de la régression hédonique entre les prix de transactions immobilières parisiennes en 2007 et leurs caractéristiques physiques. Ces 4 figures (A, B, C et D) présentent le *H-Scatterplots* pour 4 intervalles de la distance *h* (0 ; 25], (25 ; 50], (50 ; 75] et (75 ; 100]. Le nuage de points du graphique (A) correspondant aux résidus pour une distance comprise entre 0 et 25 mètres est centré sur la droite à 45°, ces résidus sont fortement corrélés. L'autocorrélation est estimée à 0,393. Lorsque la distance augmente à 25-50 mètres, le nuage de points du graphique (B) se diffuse et les points sont plus éloignés de la droite à 45° que ceux du graphique (A). L'autocorrélation diminue de 0,393 à 0,185. Les nuages de points sont de plus en plus diffus pour les graphique (C) et (D) qui appliquent des distances plus éloignées. Les

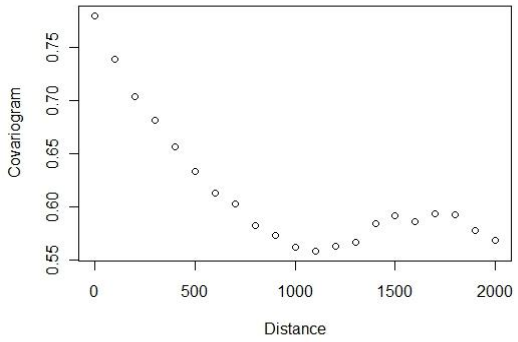
corrélations indiquées tombent à 0,0419 et 0,0827. Grâce à *H-Scatterplot,* nous observons bien que les résidus de l'estimation hédonique présentent une structure de dépendance spatiale. Plus la distance séparant les biens augmente, moins les valeurs de transaction de ces biens sont corrélées.

3.2. Covariogramme

L'analyse de l'autocorrélation spatiale se fait en deux étapes. La première étape consiste à vérifier l'existence de l'autocorrélation spatiale des observations. La seconde nécessite de décrire la structure spatiale de ce phénomène afin de pouvoir utiliser cette structure pour effectuer une estimation, une simulation ou une prédiction. L'autocorrélation spatiale est normalement décrite comme une fonction de la distance séparant les observations. Les observations proches sont fortement corrélées et lorsque les observations s'éloignent cette corrélation diminue.

Le *H-Scatterplot,* présenté dans le point 3.1 permet de vérifier si l'autocorrélation existe. Cependant, pour montrer que l'autocorrélation spatiale est fonction de la distance qui sépare les observations, nous avons besoin de plusieurs graphiques de *H-Scatterplots*. Il existe une autre méthode qui permet simultanément de vérifier l'existence de l'autocorrélation spatiale et de montrer la relation entre le niveau de covariance et la distance séparant des observations : le covariogramme. Pour chaque distance, la covariance entre les observations séparées par cette distance est estimée. L'axe des abscisses du covariogramme correspond à la distance et l'axe des ordonnés est la covariance estimée. Généralement, le covariogramme est une fonction décroissante et devient plate après une certaine distance limite. Cette distance limite, appelée la portée, est la distance à partir de laquelle il n'existe plus de corrélation entre les observations.

Figure I.4 : Le covariogramme des résidus obtenus par la régression hédonique



Les résidus obtenus de la régression linéaire entre les valeurs de transactions des appartements parisiens en 2007 et leurs caractéristiques physiques**.**

La Figure I.4 présente le covariogramme des résidus de la régression hédonique entre les prix de transactions immobilières parisiennes en 2007 et leurs caractéristiques physiques. Cette courbe décroissante du covariogramme montre que la covariance entre les observations diminue quand la distance qui sépare les observations augmente. Le covariogramme est un élément principal de la géostatistique, cette fonction est donc définie sur l'hypothèse de la continuité des observations. L'explication plus détaillée du covariogramme se trouve donc dans le CHAPITRE II, à la section 2.2.

3.3. Indice de Moran (*Moran's I*)

Une autre possibilité, celle la plus utilisée, pour vérifier l'existence d'autocorrélation spatiale entre les observations est l'indice de Moran. Comme nous l'avons indiqué, il existe deux grandes familles de méthodologies appliquées pour étudier l'autocorrélation entre les observations: la géostatistique et l'économétrie spatiale. Si le covariogramme est présenté comme un indicateur principal utilisé par l'approche géostatistique dans le cas d'une distribution aléatoire des observations dans l'espace, l'indice de Moran est un indicateur principal utilisé par l'économétrie spatiale dans le cas où les observations sont supposées distribuées de façon régulière dans l'espace.

L'indice de Moran calcule le rapport entre l'autocorrélation des variables voisines et la variance d'une telle série d'observations.

$$I_m = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{S_0} \Big/ \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n} \qquad \text{Eq. I.10}$$

avec $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$ et $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$

Le paramètre w_{ij} prend la valeur 1 si les observations y_i et y_j sont voisines et 0 sinon, S_0 qui est la somme des w_{ij} présente donc le nombre total des couples de voisinage. Le numérateur est la covariance des observations voisines. La pondération w_{ij} est appliquée pour pouvoir tenir compte uniquement de la covariance entre le voisinage. Le dénominateur de cet indice est l'espérance des carrés des écarts à la moyenne. Il représente alors la variance des observations. Remarquons que pour calculer l'indice de Moran, il faut *a priori* déterminer la structure de l'autocorrélation spatiale donc l'élément w_{ij}. Cette condition de voisinage peut être définie de plusieurs façons (l'explication détaillée se trouve dans le CHAPITRE IV à la section 2.3). Différant des deux méthodes précédentes, le calcul de cet indice présente l'inconvénient de demander une connaissance *ex ante* de la structure spatiale. Malgré cet inconvénient, son avantage important par rapport aux deux autres méthodes est que comme l'indice de Moran suit une loi normale, le test de significativité est possible.

4. Régression hédonique et problème de la dépendance spatiale

Les sections 1 à 3 servent à donner une brève information sur les données spatiales et leurs spécificités. Les données immobilières possédant de réelles caractéristiques spatiales, il est, par conséquent, impossible d'ignorer le problème de l'autocorrélation spatiale lorsque l'on travaille avec ces données. Afin de correctement évaluer la valeur immobilière, il est nécessaire de tenir compte de la distribution spatiale des observations dans le modèle d'estimation. La suite de ce chapitre consiste en un rapprochement entre

la statistique spatiale et l'étude immobilière. La régression hédonique qui est communément utilisée en évaluation immobilière et sa contrainte sont discutées. Par la suite, le problème de corrélation spatiale et ses causes sont évoqués.

4.1. Estimation hédonique des valeurs immobilières

Basu et Thibodeau (1998) donnent les détails assez complets des caractéristiques qui contribuent à donner une estimation du prix d'un bien résidentiel et qui doivent être prises en compte dans le modèle d'évaluation :

- les caractéristiques de l'immeuble (la taille de l'immeuble, la forme, la topographie, la façade, etc.),
- les caractéristiques physiques des biens (la surface en mètre carré, le nombre de pièces, le nombre de salles de bain, l'étage, le nombre de garages, la présence de piscine, les équipements, etc.),
- les caractéristiques de voisinage (le pourcentage de terrains améliorés dans le voisinage, le pourcentage de maisons occupées de propriétaire, le pourcentage de propriétés non résidentielles, le pourcentage de propriétés non développées, le temps de réponse des pompiers ou des policiers, l'indice de criminalité, etc.),
- les variables d'accessibilité (la distance au quartier central des affaires, la distance à une école, aux supermarchés, à un transport en commun, à des autoroutes importantes, etc.),
- les variables d'externalité (le bruit, la pollution, la congestion, etc.),
- les variables de zoning (la division de quartier en différentes zones : industrielle, résidentielle, loisir),
- et la date de transaction.

Dans cette étude, ces caractéristiques peuvent être regroupées en quatre catégories suivantes : **les caractéristiques physiques** qui combinent les caractéristiques de l'immeuble et les caractéristiques physiques (le type de bien, la surface, le nombre de pièce, le nombre de salle de bain, l'étage, l'existence de l'ascenseur, la qualité de l'immeuble, etc.), **les caractéristiques de localisation** qui combinent les variables de zoning et les variables d'accessibilité (l'adresse, l'arrondissement, le type de voie, la

distance jusqu'au centre d'affaires, le nombre d'écoles proches, etc.), **les caractéristiques de l'environnement** qui combine les caractéristiques de voisinage et les variables d'accessibilité (la qualité de l'environnement, la qualité des voisins, etc.) et **les catégories socioprofessionnelles et démographiques de l'acheteur** (la profession, l'âge, le revenu, le sexe, le statut patrimoniale, etc.). Remarquons que selon certaines littératures (Basu et Thibodeau (1998); Kain et Quigley (1970); Tu, Sun et Yu (2007); Wilhelmsson (2000)), uniquement les trois premiers types de caractéristiques sont mentionnés. Suite à des études plus récentes, les catégories socioprofessionnelles et démographiques de l'acheteur apparaissent aussi comme une des caractéristiques importantes (Adair, McGreal, Smyth, Cooper et Ryley (2000); Bruckel, Cusin, Juillard et Simon (2009); Macpherson et Sirmans (2001)). Afin de faciliter la distinction entre les différentes caractéristiques d'un bien immobilier, le Tableau I.1 détaille les caractéristiques prises en compte dans l'évaluation des valeurs immobilières.

Tableau I.1 : Les caractétistiques prises en compte pour estimer la valeur immobilière

Groupes de caractéristiques	Caractéristiques détaillées par Basu et Thibodeau (1998)	Exemples
Caractéristiques physiques	Caractéristiques de l'immeuble	la taille de l'immeuble, la forme, la topographie, la façade
	Caractéristiques physiques de biens	le type de bien, la surface, le nombre de pièces, le nombre de salle de bains, l'étage, l'existence d'ascenseur
Caractéristiques de localisation	Variables de zoning	l'adresse, l'arrondissement, le type de voie, la division de quartier en différentes zones : industrielle, résidentielle, loisir
	Variable d'accessibilité	la distance au quartier central des affaires, la distance à une école, à des supermarchés, à un transport en commun, à des autoroutes importantes,
Caractéristiques de l'environnement	Caractéristiques de voisinage	le pourcentage de terrains améliorés, le pourcentage de maisons occupées de propriétaire, le pourcentage de propriétés non résidentielles, le pourcentage de propriétés non développées, le temps de réponse des pompiers ou des policiers, l'indice de criminalité
	Variables d'externalité	la qualité de l'environnement, le bruit, la pollution, la congestion,
Catégories socioprofessionnelles et démographiques de l'acheteur	Néant	la profession, l'âge, le revenue, le sexe, le statut patrimoniale, le niveau de l'étude
Variable temporelle	Date de transaction	

Le modèle hédonique est une des méthodes souvent utilisées pour estimer la valeur immobilière. En régressant la valeur des biens en fonction de leurs caractéristiques, le modèle hédonique estime la valeur implicite correspondant à chaque caractéristique. La somme de toutes ces valeurs donne ainsi le prix immobilier estimé (Rosen (1974)). La régression hédonique standard pour valoriser un bien immobilier est caractérisée par l'Eq. I.11. Supposons que la valeur immobilière soit décomposée en quatre parties : celle qui correspond aux caractéristiques physiques (X_1), celle qui correspond aux caractéristiques de localisation (X_2), celle qui correspond aux caractéristiques de voisinage (X_3) et celle qui correspond aux catégories socioprofessionnelles et démographiques de l'acheteur (X_4). La régression peut être définie par :

$$y = \alpha + X_1\beta + X_2\gamma + X_3\delta + X_4\theta + \varepsilon \qquad \text{Eq. I.11}$$

y est le vecteur des prix des biens immobiliers (à la localisation S_i pour $i = 1, \dots, n$), cette valeur peut être le prix de transaction, le prix annoncé ou la valeur estimée du bien. Les coefficients β, γ , δ et θ sont les coefficients de la régression correspondant à X_1 qui est la matrice des caractéristiques physiques, X_2 qui est la matrice des caractéristiques de localisation, X_3 qui est la matrice des caractéristiques de voisinage et X_4 qui est la matrice des catégories socioprofessionnelles et démographiques de l'acheteur. α et ε sont les vecteurs des constantes et des résidus de l'estimation.

Sous la condition nécessaire que les résidus de l'estimation sont des variables aléatoires indépendantes et identiquement distribuées (la variance est constante et non corrélée), la méthode des moindres carrés ordinaires (MCO) permet d'obtenir les coefficients non biaisés de l'Eq. I.11. Cette régression hédonique standard ne permet pas de prendre en compte l'interaction existante entre les prix des biens voisins. Par contre, en réalité, la valeur d'un bien a une influence sur la valeur des autres biens situés proches de lui. Dans le cas de l'existence de corrélation spatiale entre les prix immobiliers, la méthode des moindres carrés ordinaires est insuffisante et inadaptée. Les paramètres estimés de la régression sont biaisés et non efficients et c'est la raison pour laquelle la statistique spatiale prend sa place dans l'étude immobilière. L'existence d'autocorrélation spatiale peut être signifiée par la présence de dépendance dans les résidus de l'estimation(ε). Plus précisément, l'autocorrélation spatiale apparaît quand les résidus de

l'estimation situés au point S_i sont corrélés avec ceux du point S_j. La matrice de variance-covariance, $Cov\{\varepsilon(s_i), \varepsilon(s_j)\} = \Omega$, présente des valeurs non diagonales différentes de zéro.

La dépendance spatiale peut être comparée à la dépendance temporelle. Le modèle d'estimation dans le cas de présence de dépendance temporelle des prix immobiliers est défini de façon suivante : pour chaque observation observée au point S_i, le prix observé à la date *t* dépend de celui observé aux dates antérieures, donc $y_{i,t}(y_{i,t-1}, y_{i,t-2}, \dots, y_{i,t-n})$. Le modèle de statistique spatiale prend en compte, en plus de la dimension temporelle, la dépendance spatiale. Le prix immobilier, pour chaque observation *i* située à la localisation notée S_i et pour chaque date *t*, dépend des observations situées autours de S_i, donc $y_{i,t}(y_{j,t})$. Ceci est vrai pour toutes les dates *t* si cette structure de dépendance est temporellement stationnaire.

Afin de choisir le modèle de statistique spatiale le plus approprié à l'objectif de l'étude immobilière, il faut, tout d'abord, connaitre les sources de l'autocorrélation spatiale. C'est l'objet de la suite de cette section (section 4.2). Les différentes approches de statistique spatiale utilisées dans l'étude immobilière sont détaillées dans la section 5 de ce chapitre.

4.2. Sources de l'autocorrélation spatiale

Basées sur différents travaux (Basu et Thibodeau (1998); Bowen, Mikelbank et Prestegaard (2001); Dunse et Jones (1998); LeSage et Pace (2009); Tu, Yu et Sun (2004)), trois sources d'autocorrélation spatiale sont souvent citées : la ressemblance des biens voisins, le processus de détermination des prix immobiliers et la mauvaise définition du modèle d'estimation.

Les prix des biens immobiliers sont spatialement corrélés parce que **les biens voisins sont semblables**. Cette ressemblance apparait dans les caractéristiques physiques, dans les caractéristiques de localisation et aussi dans les caractéristiques de l'environnement. Les biens voisins sont souvent construits à la même période. Par conséquent, ils ont souvent la même structure, le même style et la même taille. Les

immeubles voisins ont souvent la même qualité. Il est donc normal que la qualité d'un immeuble ait une influence sur le prix des appartements situés dans l'immeuble à côté. Tu, Sun et Yu (2007) étudient l'autocorrélation spatiale des appartements situés dans les immeubles de grand standing et indiquent que la qualité de l'espace en commun - qui est une des caractéristiques de l'immeuble - parait très importante pour évaluer la valeur de l'appartement. La dépendance des caractéristiques physiques est, par conséquent, une source d'autocorrélation spatiale des prix immobiliers. En outre, les biens voisins se confrontent normalement aux mêmes caractéristiques de localisation (la qualité de l'environnement, la présence de transports en commun, la distance jusqu'au centre ville, etc.) et aux mêmes externalités, appelées aussi les caractéristiques de l'environnement (la qualité de l'air, la vue, la nuisance sonore, etc.). Basu et Thibodeau (1998) indiquent, dans leur étude de segmentation du marché immobilier de la ville de Dallas, que les biens voisins sont confrontés à la même qualité de service municipal, à la même accessibilité et aux mêmes externalités. La non prise en compte de la dépendance des caractéristiques de localisation cause le problème de corrélation des résidus de l'estimation hédonique. La dépendance spatiale des caractéristiques sociodémographiques peut être une autre cause de dépendance spatiale des prix immobiliers. Le niveau de revenu, le niveau d'étude et la profession des propriétaires sont spatialement corrélés. Les propriétaires d'un quartier ont parfois la même profession ou le même niveau de revenu (Dubin et Sung (1990)). Dans le cas où la dépendance spatiale des prix immobiliers est causée par la structure de dépendance des caractéristiques explicatives, l'Eq. I.11 peut être modifiée afin de prendre en compte cette dépendance :

$$y = \alpha + \rho WX\beta + \varepsilon \qquad \text{Eq. I.12}$$

β est le coefficient de la régression correspondant à X qui est la matrice des caractéristiques du bien. W est la matrice de pondération permettant d'accorder des poids variés aux observations voisines. Ce poids accordé est normalement une fonction inverse de la distance séparant les biens : plus les biens sont éloignés, moindre est le poids accordé. ρ est le coefficient d'autocorrélation.

Une deuxième raison souvent employée pour expliquer la dépendance spatiale des prix immobiliers est **le processus de valorisation** du bien par le vendeur. Pour déterminer la valeur à la vente de son bien, le propriétaire peut se renseigner, soit auprès de l'expert

de quartier qui donne une estimation de prix basée sur la valeur de transaction des biens voisins, soit directement auprès de propriétaires de biens voisins. La valeur du bien immobilier dépend donc de la valeur de transaction des biens voisins. Dans le cas où la dépendance spatiale des prix immobiliers est causée par le processus de valorisation du bien, la dépendance se présente donc dans la partie des variables endogènes du modèle de régression hédonique. LeSage et Pace (2009) redéfinissent un modèle hédonique plus avancé (par rapport à l'Eq. I.11) qui permet de prendre en compte la dépendance des variables endogènes de la façon suivante :

$$y = \alpha + \beta X + \rho W y + \varepsilon \qquad \text{Eq. I.13}$$

Wy la matrice qui permet de prendre en compte la valeur à la vente des biens voisins, ρ qui permet de mesurer le degré de dépendance entre les biens voisins, βqui comporte les coefficients de la régression, X qui est la matrice des caractéristiques physiques inclues dans la régression, l'équation suppose bien qu'il n'y ait pas de dépendance spatiale des caractéristiques physiques du bien. Les vecteurs α et ε correspondent aux vecteurs des constantes et des résidus de l'estimation.

La dépendance spatiale entre les valeurs de biens immobiliers peut aussi être expliquée par les **variables omises** qui ne sont pas prises en compte dans le modèle de l'estimation. Cela peut correspondre aux variables non observables liées à la préférence individuelle (la concentration d'étrangers dans le même quartier de domicile), aux infrastructures (le projet de construction de nouvelles routes), aux nouvelles externalités (le problème d'inondation d'un certain quartier), ces informations étant manquantes dans la base de données. Par exemple, citons le cas où les valeurs de transaction de certains biens sont anormalement plus élevées que la valeur estimée juste après la date de l'annonce du classement de l'école primaire. Comme l'information concernant la qualité de l'école n'est pas disponible dans la base de données, cette information ne peut pas être intégrée dans le modèle hédonique. Cette variable manquante crée donc un problème de dépendance spatiale des prix des biens situés autour des écoles biens classées. La prise en compte de la valeur des biens voisins dans le modèle d'évaluation des biens immobiliers permet donc d'améliorer le pouvoir explicatif du modèle. Supposons que l'Eq. I.11 soit décomposée en deux parties :

$$y = \alpha + \beta X + \theta Z + \varepsilon \qquad \text{Eq. I.14}$$

avec βX et θZ représentant la valeur correspondant aux deux groupes de variables explicatives. X et Z sont indépendants mais Z subit le problème de dépendance spatiale. Si l'information sur Z est omise, l'Eq. I.11 va devenir :

$$y = \alpha + \beta X + \varepsilon \qquad \text{Eq. I.15}$$

et la dépendance se présente dans les résidus de l'estimation (ε) avec :

$$\varepsilon = \theta W \varepsilon + \epsilon \qquad \text{Eq. I.16}$$

Remarquons que la variable non observable ne cause pas uniquement un problème de corrélation spatiale, mais peut aussi générer un problème d'hétérogénéité spatiale. L'existence d'un parking privé dans un immeuble augmente le prix des appartements localisés dans cet immeuble, et l'absence d'information sur l'existence du parking cause le problème de corrélation spatiale des résidus de l'estimation de prix. En même temps, l'existence de parking privé paraît comme un caractère précieux qui augmente significativement le prix vendu de l'appartement à Paris. Par contre, le parking privé apparaît comme un caractère standard pour l'appartement en province. La valeur correspondant à ce parking varie selon la localisation du bien immobilier. L'absence d'information sur l'existence du parking, cette fois-ci, peut causer le problème d'hétérogénéité spatiale. Le choix entre ces deux problèmes de dépendance spatiale, l'autocorrélation spatiale ou l'hétérogénéité spatiale, dépend de l'interprétation de l'information et aussi de l'objectif de l'étude. Si l'étude concerne la comparaison des prix au mètre carré de parking en centre ville et en dehors du centre, le modèle avec l'hétérogénéité spatiale apparaît plus approprié. Par contre, si l'étude concerne une estimation des prix des appartements dans un immeuble de grand standing, le modèle avec l'autocorrélation spatiale paraît mieux adapté.

Ces trois sources causent le problème de dépendance spatiale qui est souvent cité dans la revue de littérature. Les chercheurs proposent à ce titre plusieurs méthodes pour le résoudre. Le point suivant permet de détailler quelques méthodes d'analyse afin d'étudier cette dépendance spatiale.

5. Analyse des données présentant une dépendance spatiale

Plusieurs méthodes sont employées pour redéfinir la régression hédonique afin d'obtenir un estimateur non biaisé et efficient de la valeur immobilière. Pace, Barry et Sirmans (1998) indiquent que les méthodes utilisées pour se confronter au problème d'autocorrélation spatiale peuvent être séparées en deux groupes : celles qui essaient de réajuster la partie des variables explicatives de l'équation (les régresseurs) et celles qui analysent les résidus de la régression. La première méthode consiste à modifier la partie des variables explicatives (βX) en ajoutant dans l'équation de régression des variables indiquant la localisation dans le but d'obtenir finalement des résidus indépendants de l'estimation. La deuxième méthode analyse les résidus de l'estimation afin d'estimer le degré de dépendance entre les observations voisines et de redéfinir le modèle de régression en incorporant ce degré de dépendance dans le modèle de l'estimation.

Ces deux méthodes ont chacune des avantages et des inconvénients. La première méthode qui travaille sur la partie des régresseurs est beaucoup plus facile à appliquer et ne nécessite pas d'information précise sur la localisation des biens, elle ne nécessite notamment pas de référence géographique. De plus, comme ce modèle ne travaille que sur la partie des variables explicatives, le temps nécessaire pour appliquer cette méthode est moins important que celui de la deuxième méthode. Cependant, afin de pouvoir éliminer totalement l'autocorrélation spatiale des résidus, il est parfois nécessaire d'inclure un nombre important de variables explicatives, donc une base de données très détaillée est indispensable. Concernant le traitement des résidus de l'estimation, cette méthode est plus compliquée. Elle nécessite les données précises de la localisation des biens et demande plus de temps pour le traitement des données. Néanmoins, cette deuxième méthode permet d'analyser le degré de dépendance entre les observations et donne une estimation plus précise.

5.1. Modélisation de la partie des régresseurs

Une des raisons pour laquelle les résidus de l'estimation sont corrélés est la mauvaise spécification du modèle de régression. Il est possible que certaines variables explicatives pourtant nécessaires soient ignorées lors de la définition de modèle. Si la dépendance spatiale est causée uniquement par ces variables manquantes, une fois ces variables spatiales ajoutées dans la partie des régresseurs, les résidus devraient être spatialement indépendants. Dans le cas de l'étude immobilière, les variables considérées comme indicateurs spatiaux sont nombreuses. Selon la revue de la littérature, certains auteurs ajoutent dans leur régression hédonique les coordonnées de localisation, la distance jusqu'au transport en commun, la distance jusqu'à l'école, la distance jusqu'au centre commerciale, le numéro de l'arrondissement, le pourcentage de l'espace vert, le nombre de parking public, etc.

Cette méthode est utilisée grâce à sa simplicité. Elle ne nécessite pas d'information précise sur les coordonnées géographiques de localisation. Cependant, il faut ajouter un nombre important de variables explicatives afin d'éliminer totalement le problème d'autocorrélation spatiale. En outre, l'application de cette méthode demande une bonne connaissance du terrain pour pouvoir définir le nombre de variables explicatives à ajouter. En effet, un nombre trop important de variables explicatives peut causer un problème de multi-colinéarité et réduire le pouvoir explicatif du modèle. La distance jusqu'à l'école risque par exemple d'être corrélée à la distance jusqu'au centre commercial parce que ces deux endroits se trouvent souvent au centre-ville. En outre, il est difficile voire quasiment impossible de définir une règle générale indiquant quelles sont les variables à ajouter dans la régression, afin d'obtenir la meilleure estimation. Par exemple, il est impossible de sélectionner parmi la distance jusqu'au centre ville, la distance jusqu'au centre commercial ou la distance jusqu'à l'école, le nombre et les variables à ajouter dans le modèle de régression afin de mieux estimer la valeur des biens immobiliers.

Néanmoins, certains auteurs mentionnent que cette méthode permet uniquement de diminuer l'autocorrélation spatiale. Plusieurs articles montrent que, malgré le nombre important de variables spatiales inclues dans la régression, une dépendance des résidus subsiste (Bourassa, Cantoni et Hoesli (2007); Pace, Barry et Sirmans (1998)). LeSage et

Pace (2009) expliquent qu'afin d'éliminer le problème de l'autocorrélation, il est nécessaire de travailler sur deux niveaux : le niveau de la structure de l'étude ainsi que le niveau de la spécification du modèle. Au niveau de la structure de l'étude, il faut prendre en compte la dimension spatiale dans le développement du modèle. Au niveau de la spécification du modèle, le choix des indicateurs de localisation est important. L'information concernant la localisation incorporée dans la base de données repose souvent sur des indicateurs de segmentation administrative comme par exemple le pays, la région, le département ou l'arrondissement. Il est possible que l'information de segmentation administrative ne soit pas suffisante pour montrer l'impact lié à la localisation. L'influence de la localisation n'est pas limitée par la segmentation administrative. Par exemple la pollution ou la nuisance sonore causée par l'implantation d'une usine n'a pas un impact négatif seulement sur la valeur immobilière dans le quartier où cette usine est implantée, mais aussi sur celles du quartier voisin. Un modèle de régression qui ne tiendrait compte que de cet indicateur administratif comme variable explicative donnerait un résultat biaisé et les résidus de l'estimation resteraient corrélés.

Toutes ces raisons incitent les chercheurs à choisir d'autres méthodes plus efficaces à appliquer. C'est la raison pour laquelle intervient la statistique spatiale qui analyse le degré de dépendance spatiale des résidus de l'estimation. Il existe deux approches permettant d'analyser les résidus : l'approche de l'économétrie spatiale et l'approche géostatistique.

5.2. Modélisation de la partie des résidus

La deuxième méthode la plus souvent appliquée pour étudier l'autocorrélation spatiale est d'analyser la partie des résidus de l'estimation. Au lieu de modifier la partie des variables explicatives de l'équation d'estimation pour trouver le meilleur modèle d'estimation présentant des résidus indépendants, cette deuxième méthode consiste à traiter directement la partie des résidus de l'estimation et à mesurer le degré de dépendance. L'objectif de cette deuxième méthode consiste premièrement à déterminer la source de l'interdépendance, puis à définir le niveau d'autocorrélation en fonction de la

distance entre les observations et finalement à essayer d'élaborer un modèle spécifique aux résidus pour retrouver les meilleures estimation et prévision.

Deux approches sont souvent évoquées pour analyser la dépendance spatiale des résidus. L'une est l'approche géostatistique qui est basée sur le principe de la géographie avec des observations distribuées dans un ensemble spatial continu. L'autre concerne l'approche de l'économétrie spatiale qui est basée sur le principe de la géométrie avec des observations distribuées comme des pixels dans un réseau discret fini. Chaque méthode est basée sur différentes hypothèses contraignantes et fournit certains avantages et inconvénients. Le choix entre ces deux méthodes dépend de l'objectif de l'étude, de la distribution des observations dans l'espace et des informations disponibles dans la base de données. Une explication détaillée sur ces deux approches est donnée dans le CHAPITRE II et le CHAPITRE IV, cette section ne donnant que des informations à titre indicatif sur chaque approche.

5.2.1. Approche géostatistique

La géostatistique a initialement été développée pour s'appliquer à l'étude de gisements miniers. Elle est donc fondée sur l'hypothèse que les observations sont distribuées continument dans l'espace. Cette approche essaie de définir directement le niveau de corrélation spatiale entre des résidus en passant par l'étude du covariogramme et du semivariogramme. Le covariogramme définit la covariance entre les observations en fonction de la distance (et aussi parfois de la direction) séparant ces observations. Ce covariogramme est décroissant avec la distance, ce qui signifie que plus la distance est importante, moins les résidus de l'estimation sont corrélés. Le semivariogramme peut être considéré comme l'inverse du covariogramme. Le semivariogramme est une fonction croissante de la distance : plus la distance est importante, plus les résidus sont dispersés et moins ils sont corrélés. Grâce à l'analyse de ces deux fonctions, la méthode géostatistique permet de définir le niveau de la dépendance spatiale en fonction de la distance et de la direction entre chaque couple des observations. Ce niveau de dépendance spatiale peut être utilisé dans la définition de la matrice variance-covariance qui est utilisée, par la suite, pour ré-estimer les paramètres du modèle de régression par les moindres carrés généralisés. De plus, la géostatique permet de définir la distance limite à partir de laquelle

la dépendance spatiale disparaît. Cette distance limite est aussi utilisée dans la segmentation des observations en plusieurs groupes selon leur niveau de corrélation.

La géostatistique définit la covariance en fonction de la distance entre les observations, des informations précises sur la localisation de chaque observation, comme les coordonnées géographiques ou les coordonnées cartésiennes, sont donc nécessaires. De plus, comme cette méthode calcule la distance et la direction entre chaque couple de biens, ce modèle est difficile à appliquer et nécessite un temps de traitement important.

5.2.2. Approche d'économétrie spatiale

L'approche issue de l'économétrie spatiale diffère de l'approche géostatistique sur plusieurs points. Premièrement, sous l'hypothèse de distribution des observations dans l'espace, l'approche issue de l'économétrie spatiale définit chaque observation comme un pixel dans un espace discret régulier. Deuxièmement, cette approche ne travaille pas directement avec la covariance, mais intègre la matrice de poids dans l'équation de l'estimation qui permet de capturer les informations concernant le degré de dépendance spatiale entre les observations voisines. L'objectif principal de cette approche est d'estimer la valeur d'une observation en prenant en compte les valeurs des observations voisines. Cette approche commence donc par la définition de l'ensemble des voisinages selon la condition de distance ou la contiguïté. Ensuite, les éléments de la matrice de poids de taille $n \times n$ (avec n le nombre d'observations) sont définis selon la condition de distance séparant chaque couple d'observations. Des poids positifs sont distribués entre les couples d'observations voisines. Ils peuvent correspondre à l'inverse de la distance, l'inverse de la distance au carré, le rapport entre l'inverse de la distance et la distance maximale dans l'ensemble des voisinages, etc. La définition des poids attribués peut être différente selon l'objectif de l'étude mais la condition principale de l'attribution des pondérations est que plus l'observation voisine se trouve lointaine, moindre est l'importance accordée à cette observation voisine.

Cette approche parait plus simple à appliquer que celle issue de la géostatistique, mais elle est basée sur certains choix arbitraires comme le choix de la condition de voisinage ou le choix du poids accordé à chaque observation voisine. De plus, si le

nombre d'observations (n) est important, la matrice de taille $n \times n$ est difficilement définie et nécessite un équipement puissant pour l'élaborer.

L'explication détaillée et la discussion sur la possibilité d'appliquer ces deux approches dans une étude immobilière sont données dans le CHAPITRE II et le CHAPITRE IV concernant l'approche géostatistique et l'approche issue de l'économétrie spatiale.

6. Littérature sur la statistique spatiale et l'étude immobilière

La statistique spatiale a été développée à la fin de 17ème siècle par Edmund Halley (Bailey et Gatrell (1995)). En étude immobilière, le regard porté sur la statistique spatiale commence par le travail de Can (1990) qui stipule que les prix des maisons voisines sont semblables seulement parce qu'ils partagent les mêmes caractéristiques de localisation. De plus, comme un agent immobilier évalue le prix d'une maison en se basant non seulement sur son emplacement, mais aussi sur les prix des maisons voisines, il est donc nécessaire d'intégrer un terme qui permette de prendre en compte cette dépendance spatiale dans la spécification du modèle. Les chercheurs commencent alors à douter de la fiabilité de l'estimation hédonique, et donc à accorder plus d'importance au problème d'autocorrélation spatiale entre les valeurs immobilières. Quelques années plus tard, la statistique spatiale se développe sur le terrain de la recherche immobilière. L'autocorrélation spatiale est mentionnée comme une contrainte dans le cas de l'évaluation immobilière (Can (1992); Des Rosiers, Thériault et Villeneuve (2000); Pace et Barry (1997)) et dans le cas de la prévision (Dubin (1998); Valente, Wu, Gelfand et Sirmans (2005)). En 1998, la revue The *Journal of Real Estate Finance and Economics* consacre un numéro spécial (volume 17 numéro 1) destiné à publier les articles concentrés sur l'analyse spatiale appliquée en finance de l'immobilier. Cette section détaille donc la revue de la littérature, selon les différentes phases de développement de l'analyse spatiale en étude immobilière.

Le développement de l'analyse spatiale dans l'étude immobilière peut être articulé en trois phases. Lors de la première phase, les chercheurs commencent à intégrer l'importance de la caractéristique de localisation dans l'évaluation des prix immobiliers. Ils incluent les indicateurs spatiaux dans la régression hédonique. Lors de la deuxième phase, grâce au développement du système d'information géographique (SIG), les données immobilières sont collectées avec leurs références géographiques. Ces références donnent une information plus précise sur la localisation des biens, ce qui permet d'appliquer l'analyse spatiale à l'étude immobilière. Et lors de la troisième phase, des modèles plus sophistiqués sont employés dans l'étude du marché immobilier comme le modèle spatio-temporel, l'analyse variogramme ou le krigeage.

6.1. Régression hédonique et caractéristiques spatiales

Parmi les facteurs déterminants de la valeur immobilière (détaillées dans 4.1), l'accessibilité et les caractéristiques de voisinage sont les deux caractéristiques spatiales auxquelles les chercheurs accordent le plus d'attention.

En ce qui concerne l'accessibilité, l'indicateur le plus souvent utilisé est la distance au centre des affaires. Le travail de Kain et Quigley (1970) est l'un de premier article qui s'intéresse aux indicateurs spatiaux. Il intègre, au sein de la régression, la distance au centre des affaires comme indicateur de l'accessibilité, mais le coefficient estimé n'est pas significatif. Il argumente que ce résultat incohérent par rapport à la réalité peut être dû à la spécification de l'équation ou à la méthode utilisée pour calculer la distance. Witte, Sumka et Homer (1979) développent le modèle hédonique de Rosen (1974) et y ajoutent, dans la partie des variables explicatives, la distance au centre des affaires. Ils montrent que cette distance est une bonne mesure de l'accessibilité. Cependant, Bender et Hwang (1985); McDonald et McMillen (1990); McMillen et McDonald (1998) argumentent que la distance au centre des affaires ou la distance au centre ville est suffisante pour mesurer l'accessibilité uniquement dans le cas où la ville est de type mono-centrique. Dans ce cas, toutes les activités économiques sont censées avoir lieu au centre ville. Par conséquent, le prix des résidences est une fonction croissante de la distance au centre ville. Par contre, cette hypothèse est critiquée par

rapport au développement d'une ville contemporaine qui est normalement du type polycentrique. McDonald et McMillen (1990) et McMillen et McDonald (1998) proposent donc d'utiliser la distance aux plusieurs centres de la ville ainsi que l'accès au moyen de transport (l'autoroute, la gare ou l'aéroport) comme les caractéristiques de localisation des biens et trouvent des coefficients significatifs. Leurs résultats confirment le développement d'une ville moderne en plusieurs segmentations. Adair, McGreal, Smyth, Cooper et Ryley (2000) utilisent le temps de déplacement à la place de la distance comme un indicateur d'accessibilité, car ils considèrent que l'accessibilité peut être indirectement mesurée par le développement du système de transport public et des nouvelles infrastructures routières qui permettent de réduire le temps de déplacement. Des Rosiers, Thériault et Villeneuve (2000) ajoutent la distance et le temps utilisé pour se déplacer au centre commercial, à l'école, au collège, à l'université et à la sortie de l'autoroute comme variables de localisation afin d'évaluer la valeur immobilière. Ils aboutissent à des résultats significatifs pour la majorité des variables. Hoesli, Thion et Watkins (1997) créent un indicateur qui peut, à la fois, mesurer la qualité de localisation et l'accessibilité. Cet indicateur est obtenu en examinant l'existence de services proches (l'école, le commerce, le transport, le bruit), et en considérant une pondération en fonction de la distance au centre ville de Bordeaux. Il en résulte que la qualité de localisation pondérée a un impact positif sur le loyer des appartements à Bordeaux.

La caractéristique de voisinage est un autre critère spatial auquel le chercheur attentionne. Plusieurs indicateurs sont ajoutés dans la régression hédonique, afin de mesurer la valeur immobilière correspondant à cette caractéristique. Kain et Quigley (1970) choisissent la qualité de l'école et le taux de criminalité comme indicateurs de qualité du voisinage. Ils trouvent un résultat cohérent et argumentent que la bonne qualité de l'école attire des propriétaires au revenu élevé qui acceptent de payer plus pour l'entretien du quartier, ce qui augmente ainsi la qualité du voisinage. Dubin et Sung (1990) étudient l'importance de la caractéristique de voisinage sur la valeur immobilière en utilisant deux catégories de variables : la qualité des services municipaux (la qualité de l'école ou le niveau de sécurité) et la catégorie socioprofessionnelle de voisinage (le niveau de revenu, le niveau de l'éducation et la profession). Selon leurs résultats, la catégorie socioprofessionnelle de voisinage apparaît plus importante que la qualité des services municipaux dans la détermination des valeurs immobilières. Lin (1993) utilise le

risque d'inondation, la pollution et la nuisance sonore pour estimer la satisfaction de voisinage dans son étude des prix immobiliers à Taiwan. Clapp (2003) et Fik, Ling et Mulligan (2003) choisissent, comme une des variables explicatives dans la régression, les références géographiques ou les indicateurs de localisation spatiale. Fik, Ling et Mulligan (2003) argumentent que l'indicateur d'accessibilité, la distance ou l'indicateur de localisation ne peuvent pas totalement prendre en compte l'influence de la localisation sur les prix des logements, car il y a un nombre indéterminable d'externalités (locales et non locales) qui influent sur le prix d'un bien situé à un endroit donné.

Selon la littérature détaillée précédemment, afin de mesurer la valeur correspondant aux caractéristiques spatiales, plusieurs variables sont choisies en fonction des différents arguments exposés. Cependant, il n'existe pas de règle générale sur le nombre de variables explicatives à ajouter. Un nombre trop important de variables peut créer un problème de multi-colinéarité. De plus, malgré le nombre important de variables ajoutées, le problème d'instabilité des paramètres estimés n'est pas toujours résolu. Pace, Barry, Clapp et Rodriquez (1998); Pace, Barry et Sirmans (1998); Valente, Wu, Gelfand et Sirmans (2005) indiquent qu'il faut un nombre élevé de variables explicatives afin de pouvoir capter toutes les valeurs liées à la localisation. Des Rosiers, Thériault et Villeneuve (2000) indiquent que l'autocorrélation spatiale existe, même si le modèle hédonique inclut comme variables explicatives les distance entre plusieurs villes. En travaillant sur les loyers de bureaux, Dunse et Jones (1998) confirment le problème d'instabilité spatiale de la valeur implicite des biens et indiquent, en outre, que la relation réciproque des caractéristiques des biens paraît réaliste.

6.2. Développement du système d'information géographique et statistique spatiale

Avec les remarques sur l'autocorrélation spatiale des valeurs immobilières et sur l'instabilité des paramètres estimés, les chercheurs commencent à s'intéresser à la statistique spatiale qui permet de prendre en compte au sein de la régression, la dépendance entre les observations spatiales. En outre, grâce au développement du système d'information géographique (GIS), les observations collectées sont référencées

sur un plan à deux dimensions. Ces références géographiques facilitent le calcul de distance entre les observations et permettent en outre d'indiquer l'ensemble des voisinages et l'ensemble des aménagements autour de chaque observation (Anselin (1998); Bible et Hsieh (1996); Can et Megbolugbe (1997); Case, Clapp, Dubin et Rodriguez (2004); Rodriguez, Sirmans et Marks (1995)). Le développement des logiciels de traitement et d'analyse des données spatiales est une autre raison qui a favorisé l'implémentation de la statistique spatiale, notamment au sein des études immobilières. Selon la littérature, la statistique spatiale est appliquée dans l'étude immobilière pour deux principaux objectifs. Premièrement, en constatant que la dépendance spatiale existe, la statistique spatiale est appliquée pour mesurer l'importance de cette dépendance parmi les régions ou les sous-marchés. Deuxièmement, la statistique spatiale est combinée à la régression hédonique pour estimer correctement la valeur de chaque caractéristique des biens immobiliers et donner la meilleure prévision de prix. Can (1992) explique que les caractéristiques spatiales influencent la valeur des prix immobiliers à deux niveaux : par l'effet de contiguïté (*adjacency effects*) et par l'effet de voisinage (*neighborhood effects*). La régression hédonique avec les caractéristiques spatiales ne mesure que l'effet de voisinage, c'est-à-dire que les prix immobiliers sont corrélés parce que les biens voisins partagent les mêmes caractéristiques de voisinage. L'effet de contiguïté représente l'influence des prix d'un bien sur le prix de voisinage. Cet effet de contiguïté n'est pas limité par la segmentation administrative. Par conséquent, les caractéristiques de voisinage et les valeurs des biens voisins doivent être prises en compte pour évaluer les prix des biens immobiliers selon le modèle autorégressif spatial. Dubin (1992) indique que, dans la plupart des cas, les coefficients de la régression hédonique correspondant aux caractéristiques de voisinage et d'accessibilité ne sont pas significatifs. Ce manque de preuve empirique de ces deux caractéristiques spatiales peut provenir de la nature multicentrique de la ville ainsi que du problème de mesure de la qualité de quartier. Il propose donc une analyse géostatistique qui permet de prendre en compte l'autocorrélation spatiale des résidus et montre que la géostatistique donne une meilleure prévision des valeurs immobilières que la régression hédonique. Gallo (2002) explique que la régression hédonique en cas de présence d'autocorrélation spatiale peut être modélisée de deux façons : soit dans la partie des variables dépendantes, c'est le cas où l'autocorrélation des prix est causée par le processus d'évaluation, soit dans les termes

d'erreur si l'autocorrélation des prix est causée par des externalités. Bowen, Mikelbank et Prestegaard (2001) citent le problème de corrélation spatiale entre les prix des biens immobiliers et détaillent le développement théorique du modèle d'économétrie spatiale qui permet de prendre en compte les prix des biens voisins dans le modèle d'estimation de la valeur d'un bien. En comparant les coefficients obtenus par la régression standard et ceux du modèle d'économétrie spatiale, ils trouvent que l'impact spatial sur le prix immobilier peut être divisé en deux composantes : l'impact au niveau de la ville qui peut être mesuré par l'indicateur de la ville, du quartier ou de la région et l'impact local qui doit être mesuré par la variable d'autocorrélation spatiale. Leurs résultats confirment donc l'argumentation donnée par Can (1992). De même, Brasington (1999) estime l'importance de la qualité de l'école et Won Kim, Phipps et Anselin (2003) mesurent l'importance la qualité de l'air sur le niveau des prix résidentiels. Ils comparent les résultats de la régression hédonique traditionnelle à ceux du modèle d'économétrie spatiale et trouvent que le modèle d'économétrie spatiale présente le meilleur pouvoir d'explicatif. L'article de Wilhelmsson (2002) est un autre article qui révèle le problème de corrélation et indique que l'ajout de l'indicateur de quartier ou de distance au centre des affaires au modèle régression ne permet pas de prendre en compte toute la dimension spatiale des données immobilières. Les variables manquantes peuvent aussi induire une dépendance spatiale et rendent les coefficients de l'estimation inefficients.

6.3. Autocorrélation spatiale, autocorrélation temporelle et géostatistique

De même que les paramètres de l'estimation hédonique varient selon la localisation, ils peuvent aussi varier temporellement. Un modèle d'économétrie spatiale du type spatio-temporel est donc appliqué à l'étude immobilière. Can et Megbolugbe (1997) indiquent que la fiabilité de l'indice des prix immobiliers est fonction de la spécification du modèle hédonique. Même si la dépendance spatiale des prix immobiliers est souvent citée dans la littérature, les indices immobiliers existants sont construits sans tenir compte de cette structure spatiale. Ils construisent donc un indice immobilier en appliquant un modèle d'économétrie spatiale qui permet d'intégrer à la fois la dépendance

spatiale entre les prix des biens voisins et la dépendance temporelle entre la valeur de transaction actuelle et celle des ventes antérieures. Finalement, ils montrent que l'indice construit avec l'économétrie spatiale est le plus précis et exact pour l'estimation des prix. Pace, Barry, Clapp et Rodriquez (1998) présentent théoriquement un modèle autorégressif spatial (variables explicatives comprenant les prix voisins) et le combinent avec un modèle autorégressif temporel (variables explicatives comprenant les prix décalés) pour produire un modèle autorégressif spatio-temporel (STAR). Ce modèle permet donc de mesurer, à la fois la dépendance spatiale et la dépendance temporelle des observations. Ils comparent les résultats obtenus par le modèle STAR avec ceux obtenus par le modèle hédonique standard. Ils montrent ainsi que le modèle STAR donne la meilleure précision de l'estimation et réduit significativement l'autocorrélation des résidus. Clapp (2003) propose une approche semi-paramétrique pour compléter le modèle STAR et Gelfand, Ecker, Knight et Sirmans (2004) appliquent le modèle STAR pour estimer l'indice local des prix immobiliers. Tu, Yu et Sun (2004) considèrent que l'autocorrélation spatiale entre les prix des bureaux dans un immeuble de grand standing est causée par deux effets : l'effet de l'immeuble appelé l'effet spatial de premier ordre (la structure spatiale apparaît parmi les biens situés au même immeuble) et l'effet de voisinage, appelé l'effet de second ordre, défini par la structure spatiale qui se trouve parmi les biens situés dans le même quartier. Ils proposent donc un modèle spatio-temporel de second ordre qui permet de séparer ces deux effets et montrent que ce modèle permet de mieux capturer la dynamique de l'indice des prix de bureaux.

Pace, Barry et Sirmans (1998) expliquent qu'il existe deux approches pour analyser l'autocorrélation spatiale : soit en modélisant la matrice de poids afin de ré-estimer le modèle de l'approche issue de l'économétrie spatiale, soit en estimant le covariogramme utilisé dans l'approche géostatistique. Selon la littérature, la géostatistique est aussi appliquée dans l'étude immobilière même si elle est moins connue que l'économétrie spatiale. Basu et Thibodeau (1998) analysent le niveau de corrélation spatiale entre les prix immobiliers dans chaque région en estimant le covariogramme sphérique des erreurs de l'estimation hédonique. Leurs résultats montrent que dans certaines régions, l'autocorrélation spatiale existe parmi les biens situés dans un rayon de 1200 mètres, mais ce résultat n'est pas valable pour toutes les régions étudiées. De plus, ils trouvent que la régression hédonique traditionnelle permet de donner une meilleure

prédiction que le *krigeage*[2] dans le cas où les observations ne sont pas corrélées. Par contre, si les observations sont corrélées, le krigeage domine le modèle hédonique. L'article de Gillen, Thibodeau et Wachter (2001) applique également le modèle géostatistique et montre l'existence de corrélation spatiale entre les résidus de l'estimation hédonique. Leur travail diffère des études précédentes dans le fait qu'ils supposent que l'autocorrélation spatiale ne dépend pas uniquement de la distance séparant des observations mais aussi de la direction (hypothèse d'anisotropie). Ils trouvent que les prix immobiliers sont plus corrélés dans la direction du centre ville. Le travail de Cano-Guervós, Chica-Olmo et Hermoso-Gutiérrez (2003) et celui de Tu, Sun et Yu (2007) utilisent tous deux l'application de la géostatistique afin de déterminer la segmentation du marché immobilier. Ils appliquent l'analyse du covariogramme pour déterminer la distance à partir de laquelle l'autocorrélation spatiale entre les prix immobiliers décline et utilisent cette distance comme un critère de sélection, afin de déterminer la segmentation de marché. Ils montrent que cette nouvelle segmentation, appelée segmentation homogène, est meilleure que la segmentation administrative.

Grâce au développement empirique et théorique des deux approches, plusieurs modèles d'estimation et de prévision existent dans la littérature. Certains auteurs essaient donc de comparer la performance entre ces différents modèles. Case, Clapp, Dubin et Rodriguez (2004) comparent les 4 modèles suivants : (1) l'estimation hédonique standard qui inclut dans la régression les références spatiales (la latitude, la longitude, ces deux références au carrée et le produit des deux références), (2) le modèle spatio-temporel avec la régression locale proposée par Clapp (2003), (3) l'analyse semivariogramme pour estimer la matrice de variance-covariance avec l'estimation du maximum de vraisemblance proposée par Dubin, Pace et Thibodeau (1999) et (4) l'estimation hédonique qui permet aux paramètres de varier selon la localisation proposée par Case. Ils concluent que le modèle de Clapp, Dubin et Case donne une meilleure performance de prévision que celle du modèle hédonique traditionnel. Cela indique l'importance d'intégrer la structure de dépendance spatiale dans le modèle. Le modèle de Dubin (modèle 3) apparaît comme le modèle le plus satisfaisant parce qu'il permet de produire des estimateurs consistants et efficients et qu'il existe des logiciels disponibles qui

[2] La méthode d'interpolation spatiale issue de la géostatistique.

facilitent l'analyse. Bourassa, Cantoni et Hoesli (2007) tentent aussi de comparer la performance de prévision des 8 méthodes suivantes : la régression hédonique simple ; la régression hédonique ajustée (la moyenne des résidus du sous-marché est ajoutée pour la prévision) ; les quatre modèles géostatistiques qui sont les analyses de résidus par les variogramme exponentiel et sphérique ainsi que celle de la version plus robuste de chacun de ces variogrammes ; et les deux modèles autorégressifs conditionnel (CAR) et simultané (SAR). Ces 8 méthodes sont comparées dans le cas d'une régression hédonique avec et sans indice de segmentation. Leur comparaison montre que le modèle autorégressif donne une plus mauvaise prévision que la régression hédonique standard. La géostatistique donne une meilleure prévision que la régression hédonique simple. Par contre, l'ajustement de la moyenne des résidus dans la prédiction, qui paraît plus simple à élaborer, donne une prévision aussi performante que la géostatistique. Les résultats de Bourassa, Cantoni et Hoesli (2007) ressemblent à ceux de Case, Clapp, Dubin et Rodriguez (2004) dans le fait que la géostatistique permet de donner une meilleure prévision que la régression hédonique et le modèle d'économétrie spatiale.

7. Conclusion

Plusieurs types de données spatiales, ainsi que plusieurs méthodes de traitement, existent et sont détaillés dans ce chapitre. Afin d'analyser les données présentant une structure spatiale -d'autocorrélation spatiale et d'hétérogénéité spatiale- les deux principales approches sont l'approche géostatistique et l'approche issue de l'économétrie spatiale. Entre ces différentes approches, le choix dépend non seulement de l'objectif de l'étude mais aussi de la distribution spatiale des données. La géostatistique est développée sous l'hypothèse de distribution continue des observations dans l'espace. Inversement, l'économétrie spatiale est développée sous l'hypothèse de distribution régulière des observations. La source de l'autocorrélation spatiale est aussi un autre critère important dans le choix de la méthode. La géostatistique analyse directement les résidus de l'estimation et suppose que l'autocorrélation spatiale se présente uniquement dans la partie des résidus. En conséquence, cette méthode considère que l'autocorrélation spatiale est causée par les variables spatiales omises lors de la spécification de l'estimation

hédonique. En revanche, l'économétrie spatiale paraît plus souple sur ce point. L'équation de l'économétrie spatiale peut être développée selon la source de l'autocorrélation, soit dans la partie des variables dépendantes, soit dans la partie des variables explicatives ou soit dans la partie des résidus. Par contre, l'économétrie spatiale nécessite un choix *ex ante* de la condition de voisinage et de la condition de dépendance. Grâce au développement de l'informatique et du système d'information géographique, ces deux approches sont appliquées dans l'étude immobilière afin d'estimer la valeur immobilière, de construire l'indice immobilier, de fournir une prévision et de déterminer la segmentation du marché.

CHAPITRE II MODELE GEOSTATISTIQUE ET ETUDE IMMOBILIERE

1. Introduction

Dans la moitié du 20ème siècle, George Matheron, un mathématicien et géologue français a développé le modèle géostatistique. Initialement, la géostatistique est utilisée principalement dans le domaine de la géologie ou de la géographie. Au fil du temps, son domaine d'application s'est élargi et elle est désormais appliquée dans plusieurs domaines – la biologie, la science sociale, l'urbanisme et même la finance – pour étudier la relation entre les observations collectées dans différentes localisations de l'espace. La géostatistique devient donc une des méthodes de la statistique spatiale les plus connues pour analyser l'autocorrélation spatiale. Puisque cette méthode était destinée historiquement à l'étude de l'évolution de gisements miniers, la géostatistique de base s'est développée sous certaines hypothèses spécifiques: la continuité et la stationnarité de la distribution spatiale des observations. Ces deux hypothèses ne posent aucun problème si les observations spatiales sont des gisements de matières premières ou des concentrations de polluants le long d'une rivière. Par contre si les observations sont les prix de biens immobiliers, on peut se demander si la distribution spatiale des prix de biens immobiliers est continue et stationnaire. Selon l'hypothèse d'isotropie, il serait aussi intéressant de vérifier si l'autocorrélation entre les prix de biens immobiliers dépend uniquement de la distance séparant leurs localisations. De même, se pourrait-il que l'autocorrélation varie selon la direction, et que par exemple l'autocorrélation spatiale soit plus forte dans la direction du centre-ville ?

Remarquons que des méthodes géostatistiques plus sophistiquées se sont développées sous des conditions plus souples comme la géostatistique non-linéaire ou la géostatistique non-stationnaire, ces méthodes sont souvent utilisées dans le domaine de la science de la terre ou la science naturelle. Cependant, l'unique méthode géostatistique appliquée en étude immobilière est la géostatistique linéaire. Par conséquent, le terme géostatistique utilisé par la suite dans cette thèse se rapportera uniquement à la géostatistique linéaire.

L'analyse géostatistique se réalise en deux étapes principales. La première étape concerne l'identification de la structure spatiale des observations. Les deux principaux outils utilisés sont le covariogramme, qui permet de déterminer la variation de

l'autocorrélation mesurée entre deux points quand la distance augmente et le semivariogramme, qui décrit la variabilité des observations en fonction de la distance. Or, le variogramme empirique ne pouvant être utilisé tel quel, il est nécessaire d'ajuster un modèle de variogramme paramétrique sur le variogramme empirique afin d'en estimer les paramètres : le palier, la portée et la pépite, ces indicateurs étant très importants en géostatistique. La littérature mentionne plusieurs modèles théoriques de variogramme: le modèle linéaire, le modèle sphérique, le modèle gaussien et le modèle exponentiel. Le choix entre ces différents modèles dépend de la nature de la variable étudiée et de l'objectif de l'étude. Il paraît donc important d'identifier le modèle le plus adapté aux données immobilières et de connaître l'impact du choix ex-ante du modèle variogramme théorique sur le résultat de l'étude immobilière.

L'objectif de ce chapitre est double. Dans un premier temps, nous discuterons les deux principales hypothèses du modèle et les différents éléments nécessaires pour estimer l'autocorrélation spatiale par la géostatistique. Ensuite, nous verrons l'application de la géostatistique à l'étude immobilière. Après cette introduction, la deuxième section de ce chapitre explique les hypothèses faites, les deux outils principaux de l'étude géostatistique, l'estimation paramétrique du variogramme et cette section se termine par l'explication très brève de la prévision par la méthode du krigeage. La troisième section correspond à un rapprochement de la géostatistique et de l'étude immobilière. Les champs d'application de la géostatistique en étude immobilière sont multiples. On peut par exemple citer l'estimation de prix, la prévision de valeur, la construction d'indice, la segmentation de marché ou l'étude de l'impact sur le prix d'une nouvelle infrastructure. La revue de littérature que nous ferons par la suite sera organisée suivant les champs d'application tel que décrits ci-dessus. Afin de pouvoir répondre aux questions posées dans les paragraphes précédents, nous analyserons l'application aux données immobilières sur plusieurs points, à savoir la distribution spatiale des données immobilières, les hypothèses contraignantes de la méthode et le choix du variogramme théorique. Finalement, les avantages et les inconvénients de l'étude géostatistique dans le cas de l'étude immobilière sont présentés pour terminer le chapitre.

2. Géostatistique

La géostatistique est une étude de variables numériques réparties dans l'espace. Cette méthode est un rapprochement des deux domaines : la géographie qui est une étude de la terre d'une part, et d'autre part la mathématique. Chauvet (1999) décrit l'évolution de la géostatistique en trois périodes, appelées les trois âges de la géostatistique. La première période correspond aux années 50 : ce sont les problèmes rencontrés dans les mines d'or d'Afrique du Sud qui suscitent les premières recherches en géostatistique ; il s'agissait de mesurer la variabilité de la teneur du minerai d'or. Krige (1951) développera la méthode d'estimation pour l'industrie minière. La deuxième période, correspond aux années 60. Matheron (1965) détaille le développement théorique et mathématique de la géostatistique dans l'ouvrage « *Les variables régionalisées et leur estimation* » ou « *Principles of geostatistics* » en anglais. Cette méthode devient de plus en plus connue et s'est répandue rapidement dans d'autres domaines des sciences de la terre. Le préfixe « *géo* » de la géostatistique établit clairement un lien entre la géostatistique et la science de la terre et explique bien l'origine du développement de la méthode. La troisième période correspond à la phase d'expansion de la géostatistique à partir des années 80. Ses champs d'application ne se limitent plus désormais aux ressources naturelles. Cette méthode est développée et appliquée maintenant à un grand nombre de disciplines autre que la science de la terre ou la science naturelle. Elle gagne le terrain de la recherche en science environnementale et en science humaine et sociale ; elle est appliquée, par exemple, dans la recherche de d'infection en épidémiologie, le changement démographique et même dans le choix de l'individu en immobilier.

2.1. Hypothèses

Dans le CHAPITRE I, les données spatiales sont classées en trois catégories, la première catégorie correspond au cas où les indicateurs spatiaux sont répartis de façon aléatoire dans un espace continu. Ces sont « *les variables régionalisées* » ou « *les variables géostatistiques* ». Les notions de fonction aléatoire et de variable régionalisée jouent un rôle fondamental en géostatistique. L'hypothèse de la continuité – la distribution continue des indicateurs spatiaux – est considère comme la première

hypothèse principale de l'étude géostatistique. Afin de pouvoir appliquer l'analyse variogramme aux données réelles, il est nécessaire de supposer que le processus spatial est stationnaire. La deuxième hypothèse nécessaire est donc la stationnarité. Pour simplifier l'analyse géostatistique, l'étude immobilière impose souvent l'hypothèse d'isotropie au variogramme : l'autocorrélation spatiale ne doit dépendre que de la distance et est donc indépendante de la direction.

Les trois hypothèses fondamentales de la géostatistique – la continuité, la stationnarité et l'isotropie – sont détaillées dans cette section.

2.1.1. Continuité

Le point principal qui différencie le modèle géostatistique du modèle d'économétrie spatiale est son champ d'application. La géostatistique suppose que les indicateurs de localisation d'observations se distribuent de façon *aléatoire* dans l'espace *continu* à deux dimensions. Par contre, le modèle d'économétrie spatiale s'applique mieux dans le cas où les observations sont indexées dans un plan de réseau régulier. Cressie (1991) par ailleurs indique que la géostatistique correspond au cas où le processus spatial est *continu*. Formellement, les réalisations de variable $\{z(s_1), z(s_2), \dots, z(s_n)\}$ observées aux indicateurs de localisation $\{s_1, s_2, \dots, s_n\}$ sont distribuées de façon *continu* dans l'espace R^d où d est la dimension de l'espace de réalisation. La géostatistique suppose que le champ d'étude dans laquelle les estimations sont calculées est au moins continu mais la structure spatiale peut ne pas être homogène ou être non stationnaire. Dans ce cas, le champ d'étude ne devrait pas être coupé en deux par une frontière ou toute autre barrière qui introduit une discontinuité. S'il y a une discontinuité, les observations devraient être divisées en différents sous-ensembles distincts et l'analyse géostatistique se réalise séparément dans chaque sous ensemble.

Cette hypothèse de continuité est importante afin de pouvoir avancer les différentes étapes de la méthode géostatistique à savoir l'analyse de variogramme et le *krigeage*. En effet, la continuité spatiale se traduit par le fait que deux observations collectées en deux points ont tendance à être d'autant plus semblables que ces deux points sont proches. Par conséquent, cela conduit à une analyse de variogramme considérant

l'autocorrélation spatiale comme une fonction continue décroissante de la distance séparant les observations. D'ailleurs, l'hypothèse de continuité spatiale est aussi nécessaire afin d'établir la prévision par le krigeage. Si l'hypothèse de la continuité spatiale est faite, une observation non échantillonnée peut être estimée à partir de ses voisins. Inversement, en l'absence de continuité, une observation peut prendre n'importe quelle valeur et toute prévision est donc impossible.

2.1.2. Stationnarité

Dans la plupart des cas d'études géostatistiques, l'analyse variogramme est réalisée sous la condition que le variogramme ne varie pas selon la localisation. Par conséquent, les observations sont supposées être distribuées selon un processus *stationnaire*. Si l'hypothèse de continuité spatiale joue un rôle fondamental, l'hypothèse de stationnarité spatiale joue aussi un rôle prépondérant afin de rendre possible l'inférence géostatistique. Un processus est stationnaire si la loi de probabilité de la fonction aléatoire est invariante par translation ; elle ne dépend pas de la localisation. La *stationnarité au sens strict* nécessite que tous les moments soient invariants par translation. Néanmoins, cette stationnarité au sens strict est en pratique difficilement respectée et il est impossible d'estimer tous les moments avec le nombre limité d'observations (Armstrong et Carignan (1997)). La stationnarité au sens strict est donc remplacée par la stationnarité au sens faible, aussi appelée la stationnarité de second ordre.

Un processus est *stationnaire d'ordre 2* si les deux premiers moments sont invariants pas translation, ils ne dépendent pas de la localisation. L'espérance et la variance sont constantes en tout point de l'espace étudiée et la covariance est constante pour tous les couples de points équidistants. La stationnarité au sens faible se traduit par :

$$\forall s_i, s_j, h \in R^d$$

$$E\{z(s_i)\} = E\{z(s_j)\} = E\{z(s_i + h)\} = \mu \qquad \text{Eq. II.1}$$

$$Cov\{z(s_i), z(s_j)\} = C(s_i - s_j) = C(h) \text{ avec } h = s_i - s_j \qquad \text{Eq. II.2}$$

$$C(0) = Var\{z(s_i)\} = Var\{z(s_j)\} = \sigma^2 \qquad \text{Eq. II.3}$$

s_i et s_j sont les deux localisations voisines qui sont séparées par le vecteur de distance *h*, $E\{..\}$, $Var\{..\}$ et $Cov\{..\}$ représentent l'espérance, la variance et la covariance entre les observations. $C(h)$ représente le covariogramme qui décrit la covariance en fonction du vecteur de distance *h*.

Un processus qui satisfait les trois conditions (Eq. II.1 à Eq. II.3) est dit stationnaire au second ordre. La première condition (Eq. II.1) signifie que l'espérance existe, est constante au travers de l'espace étudié et est indépendante de la localisation. La deuxième condition (Eq. II.2) équivaut à dire que pour chaque couple d'observations $\{z(s_i), z(s_j)\}$, la covariance entre ces deux observations existe et ne dépend que du vecteur de distance *h* séparant ces deux points. La localisation de ces deux points ne joue donc aucun rôle sur la covariance étudiée. Avec l'hypothèse de stationnarité de second ordre, la covariance peut être définie comme une fonction d'un seul argument qui est le vecteur de distance et cette fonction, notée $C(h)$, est appelée *covariogramme*. La stationnarité de la covariance implique aussi la stationnarité de la variance. En effet, si $h = 0$, l' Eq. II.3 correspond à la troisième condition sur la variance. La troisième condition (Eq. II.3) signifie que la variance existe et est indépendante de la localisation. Un processus qui ne vérifie pas l'une de ces trois conditions est non stationnaire.

En pratique, il arrive souvent que ces conditions ne soient pas satisfaites. Afin d'appliquer l'analyse géostatistique, il faut, tout d'abord, transformer le processus non stationnaire en processus stationnaire. Toutefois, il est nécessaire et souvent difficile de déterminer l'origine de la non stationnarité. La source de non stationnarité observée le plus fréquemment en étudiant l'autocorrélation spatiale est l'existence d'une tendance (*trend*), appelé la non stationnarité déterministe. L'espérance varie dans l'espace, autrement dit, l'espérance varie selon la localisation ; la première condition du moment d'ordre 1 (Eq. II.1) n'est donc pas satisfaite. Ce processus peut être rendu stationnaire en lui enlevant sa tendance déterministe (*data detrending*), en divisant le champ d'étude en plusieurs sous espace (data *subsampling*) ou en transformant les observations afin d'avoir un processus stationnaire (*data transforming*) (Atkinson et Lloyd (2007)).

Néanmoins, même lorsque l'espérance est constante, la covariance n'existe pas nécessairement (Armstrong et Carignan (1997)). La stationnarité de second ordre est trop

restrictive et difficilement vérifiable dans la pratique. Une condition plus faible, plus simple et plus facile à vérifier a été proposé par Matheron (1963) correspondant à l'hypothèse de la stationnarité intrinsèque. Ainsi, afin de mettre en œuvre la géostatistique, l'hypothèse de stationnarité intrinsèque doit au minimum être remplie. Sous l'hypothèse de la stationnarité intrinsèque, seuls les accroissements du processus doivent être stationnaires d'ordre 2. Le processus spatial doit donc vérifier les deux conditions suivantes :

$$E\{z(s_i) - z(s_j)\} = 0 \qquad \text{Eq. II.4}$$

$$Var\{z(s_i) - z(s_j)\} = 2\gamma(s_i - s_j) = 2\gamma(h) \text{ avec } h = s_i - s_j \qquad \text{Eq. II.5}$$

La première condition (Eq. II.4) signifie que l'espérance des variations existe et ne varie pas au travers de l'espace étudié. La deuxième condition (Eq. II.5) montre que la variance des variations existe et est indépendante de la localisation. Le *semivariogramme*, noté par $\gamma(h)$, permet de décrire la moitié de la variance des variations en fonction du vecteur de distance *h*.

Remarquons qu'il existe une méthode géostatistique plus sophistiquée destinée à analyser une structure spatiale non stationnaire. Néanmoins, seule la géostatistique linéaire sous hypothèse de stationnarité est utilisée en étude immobilière, les autres méthodes ne seront pas traitées dans cette thèse.

2.1.3. Isotropie

Sous l'hypothèse de stationnarité, la covariance entre les observations ne dépend que le vecteur de distance (*h*). Ce vecteur contient de l'information sur la longueur – indiquée par sa norme – ainsi que sur l'orientation de *h*. En fait, si le covariogramme ne dépend que de la longueur (la norme de *h*, $\|h\|$), il est dit isotrope mais s'il dépend aussi de l'orientation du vecteur de distance, il est dit anisotrope. Autrement dit, pour une longueur $\|h\|$ donnée, la covariance estimée peut varier selon l'orientation du vecteur h (le cas d'anisotropie) ou elle peut être identique pour toutes les directions et invariante par rapport à une direction particulière (le cas d'isotropie). Sous l'hypothèse d'isotropie du

covariogramme, la covariance entre des observations dépend de la longueur séparant les observations mais pas de la direction. Il existe donc un seul et unique covariogramme qui est défini comme omnidirectionnel. Néanmoins, cette hypothèse est souvent critiquée comme étant irréaliste, surtout dans le cas de l'étude immobilière. Si la covariance est directionnellement dépendante, le processus n'est plus isotrope mais anisotrope. Le covariogramme varie selon la direction ; par exemple, pour une même distance séparant les observations, la covariance estimée pour la direction nord-sud est moins importante que celle estimée pour la direction est-ouest. Dans ce cas, plusieurs covariogrammes doivent être estimés en fonction de la direction prise en compte dans l'étude.

Les hypothèses de continuité, de stationnarité et d'isotropie de la géostatistique, ont souvent été utilisées sur des données immobilières. Plusieurs questions se posent alors ; Est-il raisonnable de considérer que les données immobilières sont distribuées continument dans l'espace sans tenir compte des frontières administratives ? Est-il raisonnable de supposer que la structure spatiale des données immobilières est stationnaire, que les prix de biens immobiliers ont le même comportement de dépendance spatiale s'ils se localisent au centre-ville ou en dehors de la ville ? Est-il raisonnable d'imposer que la dépendance spatiale des valeurs immobilières dépend uniquement de la distance séparant leurs localisations ? La direction a-t-elle de l'importance ? La section 3.2 va nous permettre de répondre à ces questions.

2.2. Covariogramme et semivariogramme

L'étude de la structure spatiale par la géostatistique est abordée par l'intermédiaire du calcul de covariance et variance. Sous les hypothèses de continuité, de stationnarité de second ordre et d'isotropie du processus spatial, la covariance et la variance peuvent être définies comme des fonctions d'un seul argument qui est la distance séparant les observations. Ce sont respectivement le covariogramme et le semivariogramme empiriques ; les deux outils fondamentaux de l'analyse géostatistique. Les sections suivantes détaillent le calcul de ces deux variogrammes ainsi que leurs propriétés.

2.2.1. Covariogramme

Définissons l'indicateur de localisation dans un plan (x, y) par $s_i = (x_i, y_i)$ et l'indicateur de localisation éloignée de la distance h par rapport au point s_i par $s_i + h$. Les observations $z(s_i)$ et $z(s_i + h)$ sont collectées à partir de ces deux localisations. La covariance entre les observations des deux points séparés par la distance *h* est définie par l'Eq. II.6.

$$Cov\{z(s_i), z(s_i + h)\} = E\{[z(s_i) - m(s_i)] \times [z(s_i + h) - m(s_i + h)]\} \quad \text{Eq. II.6}$$

avec $m(s_i) = E\{z(s_i)\}$ et $m(s_i + h) = E\{z(s_i + h)\}$ les espérances respectives des observations collectées aux localisations s_i et $s_i + h$.

Sous l'hypothèse de stationnarité de second ordre et sous l'hypothèse de d'isotropie, la covariance est la même pour tous les couples de points équidistants et cette covariance dépend uniquement de la distance séparant les observations, et non pas de la direction. En conséquence, pour toute variation de distance *h*, la covariance peut être estimée. L'ensemble des covariances estimées pour toutes les variations de distance possibles donne le covariogramme. Le covariogramme $C(\|h\|)$ définissant la ressemblance entre les observations en fonction de la distance qui les sépare est décrit de la façon suivante :

$$Cov\{z(s_i), z(s_i + h)\} = C(\|h\|) \equiv C(h) \quad \text{Eq. II.7}$$

Pour simplifier la notation, le covariogramme correspondant à la distance $\|h\|$ est noté $C(h)$. Empiriquement, à partir des observations collectées, le covariogramme est calculé par :

$$\hat{C}(h) = \frac{1}{N_h} \sum_{N_h} (z(s_i) - \mu)(z(s_i + h) - \mu) \quad \text{Eq. II.8}$$

N_h correspond au nombre de paires d'observations séparées par la distances h ; $\mu = \frac{1}{n}\sum_{j=1}^{n} z(s_i) = \frac{1}{m}\sum_{j=1}^{m} z(s_i + h)$ correspond à la moyenne des échantillons. Sous l'hypothèse de stationnarité du second ordre cette espérance est constante dans l'espace étudié.

Le covariogramme possède les propriétés suivantes :

i) la valeur de la covariance pour les observations ayant la même localisation $(h = 0)$ est égale à la variance : $Cov\{z(s), z(s)\} = Var\{z(s)\} = \sigma^2$. Par conséquent, le covariogramme correspondant à la distance zéro est égale à la variance : $C(h = 0) = C(0) = \sigma^2$,

ii) sous l'hypothèse de continuité et de régularité du champ aléatoire, le covariogramme est une fonction continue,

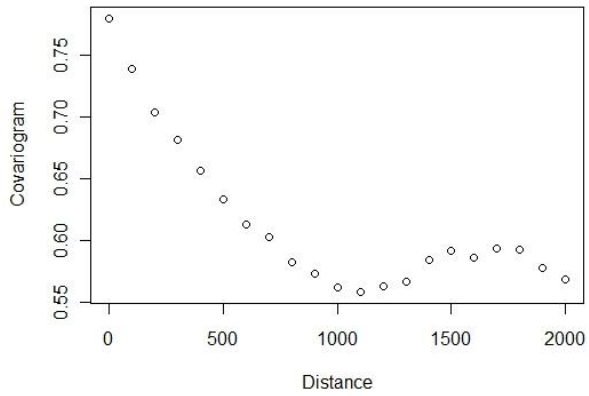
iii) le covariogramme est une fonction symétrique : $C(-h) = C(h)$,

iv) le covariogramme est une fonction décroissante de la distance ; plus la distance séparant les observations est importante, moins les valeurs observées sont corrélées : $C'(h) < 0$,

v) il existe une distance $\theta_2 > 0$, telle que lorsque $h \geq \theta_2$, $C(h) = 0$. La valeur de θ_2 vérifiant cette condition s'appelle la portée (*range*) du covariogramme. La portée est donc la distance au-delà de laquelle le covariogramme devient nul, ou encore la distance au-delà de laquelle la covariance entre les observations tend à disparaître,

vi) dans le cas où le covariogramme tend seulement asymptotiquement vers 0 quand $h \to \infty$ ($\lim_{h\to\infty} C(h) \to 0$), il est possible qu'il n'existe pas une telle portée finie, la « portée pratique » (*practical range*) est définie comme une distance $\theta_2 > 0$, telle que lorsque $h \geq \theta_2$, $C(h) \leq 0{,}05C(0)$. La portée est donc la distance au-delà de laquelle le covariogramme est réduit de plus que 95%, ou encore la distance au-delà de laquelle la covariance entre les observations est négligeable.

Figure II.1 : Le covariogramme empirique des résidus de la régression hédonique des prix résidentiels parisiens en 2007.



2.2.2. Semivariogramme

Une autre approche permettant de décrire la structure spatiale entre les observations dans l'espace est le semivariogramme ($\gamma(h)$) qui décrit la variance des variations de valeurs séparés par la distance *h*. A l'inverse du covariogramme qui mesure la ressemblance entre les valeurs prises sur deux localisations en fonction de la distance qui les sépare, le semivariogramme mesure plutôt la dissemblance entre elles. Le semivariogramme est défini par :

$$\gamma(h) = \frac{1}{2} Var\{z(s_i) - z(s_i + h)\} \\ = E\left\{\Big(\big(z(s_i) - z(s_i + h)\big) - \big(m(s_i) - m(s_i + h)\big)\Big)^2\right\} \quad \text{Eq. II.9}$$

Sous l'hypothèse de stationnarité intrinsèque, $E\{m(s_i + h) - m(s_i)\} = 0$, l'Eq. II.9 devient :

$$\gamma(h) = \frac{1}{2} Var\{z(s_i) - z(s_i + h)\} = E\{(z(s_i) - z(s_i + h))^2\} \quad \text{Eq. II.10}$$

Empiriquement, à partir des données spatiales, le semivariogramme est estimé par l'espérance des différences des valeurs observées au carrée :

$$\hat{\gamma}(h) = \frac{1}{2N_h} \sum_{N_h} [z(s_i) - z(s_i + h)]^2 \quad \text{Eq. II.11}$$

avec N_h correspondant au nombre de paires des observations séparées par la distance h, $z(s_i)$ et $z(s_i + h)$ sont les observations aux points s_i et $s_i + h$.

Les propriétés principales du semivariogramme sont les suivantes :

i) sous l'hypothèse de stationnarité du second ordre, le semivariogramme est relié au covariogramme par :

$$\begin{aligned}\gamma(h) &= \frac{1}{2}E\{(z(s_i) - z(s_i + h))^2\} \\ &= \frac{1}{2}E\{z(s_i)^2 - 2z(s_i)z(s_i + h) + z(s_i + h)^2\} \\ &= \frac{1}{2}[Var\{z(s_i)\} - 2Cov\{z(s_i), z(s_i + h)\} + Var\{z(s_i + h)\}] \\ &= \frac{1}{2}[2C(0) - 2C(h)] \\ \gamma(h) &= C(0) - C(h)\end{aligned} \qquad \text{Eq. II.12}$$

ii) le semivariogramme à la distance 0 est égale à 0 : $\gamma(h = 0) = 0$ car $z(s_i) - z(s_i) = 0$,

iii) le semivariogramme est une fonction symétrique : $\gamma(-h) = \gamma(h)$,

iv) le semivariogramme est une fonction croissante de la distance ; plus la distance séparant des observations est importante, plus les valeurs observées deviennent dispersées : $\gamma'(h) < 0$,

v) à partir d'une certaine distance limite, le semivariogramme atteint un pallier asymptotique qui est égal à la variance de l'échantillon : $\lim_{h\to\infty}\gamma(h) = C(0) - \lim_{h\to\infty}C(h) = C(0) = \sigma^2$ car $\lim_{h\to\infty}C(h) = 0$

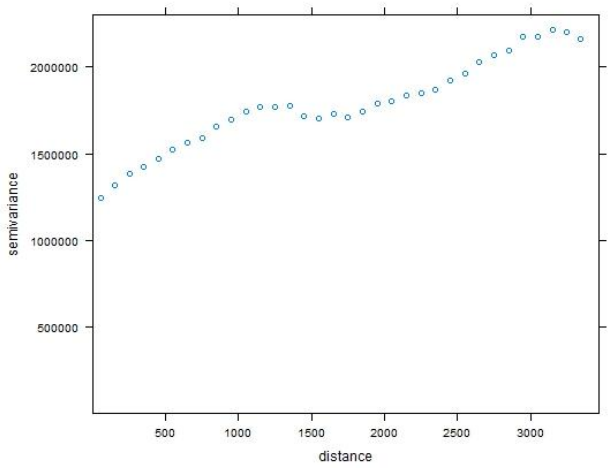
vi) il existe une distance $\theta_2 > 0$, au-delà de laquelle le semivariogramme devient stable : lorsque $h \geq \theta_2$, $\gamma(h) = \theta_1$. La valeur de θ_2 vérifiant cette condition s'appelle la portée (*range*) du semivariogramme et la valeur θ_1 est le pallier (*sill*) du semivariogramme,

vii) dans le cas où le semivariogramme tend uniquement asymptotiquement vers θ_1 ($\lim_{h\to\infty}\gamma(h) \to \theta_1$), il est possible qu'il n'existe pas de portée finie. La « portée pratique » (*practical range*) est définie comme une distance $\theta_2 > 0$, telle que lorsque $h \geq \theta_2$, $\gamma(h) \geq 0{,}95\theta_1$. La portée

pratique est donc la distance au-delà de laquelle le semivariogramme atteint plus que 95% de la valeur du pallier,

viii) il est possible que le semivariogramme soit discontinu près de l'origine : $\lim_{h\to 0}\gamma(h) = \theta_0 > 0$. Ce saut à l'origine est appelé l'effet de pépite (*nugget effect*) et le paramètre θ_0 est parfois appelé le « *nugget* ».

Figure II.2: Le semivariogramme empirique des résidus de la régression hédonique des prix résidentiels parisiens en 2007.



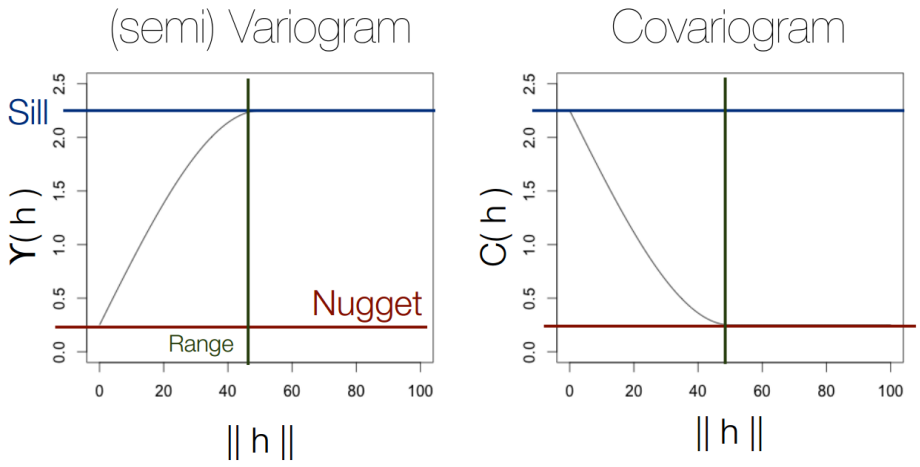
Le semivariogramme isotrope est borné et caractérisé par les trois importants paramètres suivants :

- le semivariogramme atteint un pallier constant au-delà d'une certaine distance: $\lim_{h\to\infty}\gamma(h) = \theta_1$, θ_1 est appelé le pallier ou « *sill* »,
- la distance à partir de laquelle le semivariogramme atteint quasiment son pallier (le paramètre $\theta_2 : \gamma(\theta_2) \to \theta_1$) s'appelle la portée ou « *range* » . La portée ou une « *range* » est la distance au-delà de laquelle la covariance est négligeable,
- le semivariogramme peut être discontinue près de l'origine : $\lim_{h\to 0}\gamma(h) = \theta_0$, le paramètre θ_0 est appelé la pépite ou « *nugget* » ; remarquons que lorsque $h = 0$, $\gamma(0) = 0$ par définition mais lorsque $h = 0^+$, $\gamma(h) = \theta_0$.

La littérature indique souvent que le pallier correspond à la variance des échantillons (Journel et Huijbregts (1991)). Par contre, pour Barnes (1991) cette condition n'est possible que si la distance maximale définie pour estimer le semivariogramme est

largement supérieure à la portée du semivariogramme. Si c'est le cas, le nombre de paires d'observations séparées par une distance supérieure à la portée est significativement plus élevé que le nombre de paires séparées par une distance inférieure à la portée. En conséquence, la variance de l'échantillon peut être définie comme la pallier du semivariogramme ($\theta_1 = \sigma^2$). Dans l'autre situation possible, si l'échantillon est obtenu à partir d'un espace de dimension inférieure ou égale à la portée, il n'est pas raisonnable d'estimer le pallier du semivariogramme par la variance des échantillons.

Figure II.3 : Les propriétés du semivariogramme et covariogramme



(*source : package gstat du logiciel R*)

En théorie, le semivariogramme et le covariogramme sont les deux outils permettant d'estimer la dépendance spatiale des observations. En pratique, le semivariogramme est préférable au covariogramme. Deux raisons peuvent justifier ce choix. Premièrement, le semivariogramme est basé sur des hypothèses moins restrictives que le covariogramme. Le semivariogramme ne se définit pas directement à partir de la valeur observée $\big(z(s_i)\big)$ mais plutôt à partir de la différence de valeurs $\big(z(s_i) - z(s_i + h)\big)$. L'hypothèse de stationnarité faite sur les variations est plus faible étant donné que seule la moyenne et la variance des variations des valeurs doivent être invariantes à travers l'espace étudiée. Cette hypothèse de stationnarité des variations est parfois appelée stationnarité intrinsèque. L'hypothèse de stationnarité de second ordre faite sur le covariogramme est donc plus forte que l'hypothèse de stationnarité intrinsèque faite sur le

semivariogramme. Deuxièmement, le calcul du covariogramme nécessite l'estimation de l'espérance inconnue de l'échantillon $\big(m(s_i)\big)$. Même si cette espérance est supposée constante sous l'hypothèse de stationnarité, son estimation peut induire un biais.

Remarquons qu'en pratique, il est possible d'estimer le semivariogramme empirique de façon continue. Une tolérance est donc appliquée pour déterminer un intervalle de distance (par exemple [0-50m], [50m-100m], [100m-150m], …). Il n'y a pas de règle générale dans le choix de l'amplitude de l'intervalle ; tout dépend de la taille des données et de l'objectif de l'étude. Dans le cas où le nombre d'observations est important l'intervalle de distance choisi peut être petit de façon à permettre d'obtenir une courbe du semivariogramme lisse. Cependant, si l'intervalle est trop petit, le nombre trop faible d'observations impliquées dans chaque intervalle peut causer une fluctuation des écarts de valeurs par la présence de certaines valeurs extrêmes. Dans le cas d'une taille modeste d'observations, l'amplitude choisie doit être assez large pour qu'il y ait un nombre suffisant d'observations dans chaque intervalle ; mais cela peut créer des sauts dans le variogramme empirique estimé.

2.2.3. Covariogramme et semivariogramme anisotropie

Les points sections 2.2.1 et 2.2.2 ont présenté le covariogramme et le semivariogramme dans le cas où le processus spatial est isotrope. Le variogramme (le semivariogramme et le covariogramme) dépend uniquement de la distance h séparant deux points dans l'espace. En supprimant cette hypothèse, le processus spatial est anisotrope, le variogramme dépend à la fois la distance $\|h\|$ et de l'orientation du vecteur h. Lorsque le processus est anisotrope, toutes les directions ne sont pas équiprobables, l'unique estimation du variogramme ne suffisant pas pour déterminer la dépendance spatiale, plusieurs variogrammes doivent être estimés en fonction de la direction du vecteur de distance. Deux types d'anisotropie sont souvent mentionnées dans la littérature (Cressie (1991)) ; l'anisotropie géométrique et l'anisotropie zonale. Dans le cas de l'anisotrope géométrique, le pallier est identique dans toutes les directions mais la portée diffère d'une direction à l'autre. Cela veut dire que l'autocorrélation spatiale entre les observations disparaît si la distance séparant ces observations est supérieure à une

certaine distance limite, mais cette distance limite varie selon la direction. Il est possible de corriger cette anisotropie géométrique en transformant le système des coordonnées des observations afin de pouvoir obtenir un variogramme isotrope. Dans l'autre cas, si le pallier et la portée ou même la pépite de variogramme varient selon la direction, le processus est anisotrope zonale. Il n'est pas alors possible de transformer une telle structure pour obtenir un variogramme isotrope. Cependant, la combinaison entre un variogramme isotrope et un variogramme anisotrope géométrique est parfois utilisée pour approximer le variogramme anisotrope zonal.

2.3. Estimation paramétrique de variogramme

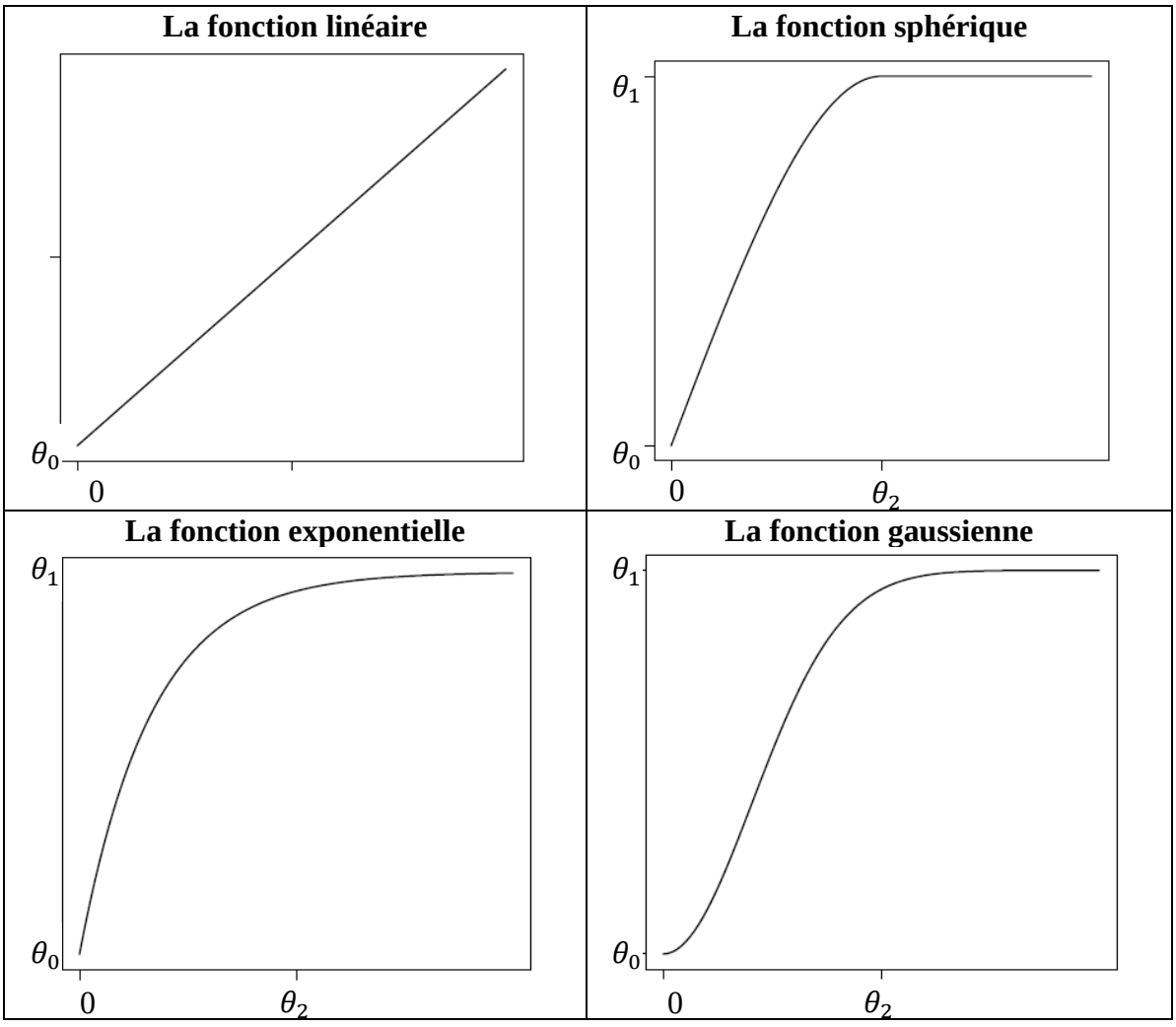
Les fonctions de variogramme empiriques présentées dans la section précédente ne varient pas nécessairement de façon régulière avec la distance parce que certaines sont estimées avec plus ou moins d'observations que les autres. Par contre, la méthode de prévision exige que l'information sur toutes les distances possibles soit disponible. Ainsi donc, le covariogramme et le semivariogramme devront être des courbes continues et lisses. Une fois le covariogramme et le semivariogramme calculés, il faut ensuite ajuster un modèle paramétrique de variogramme sur le variogramme empirique afin d'estimer la portée, le pallier et la pépite.

L'ajustement du variogramme paramétrique sur le variogramme empirique se déroule en deux étapes. Dans une première étape, il faut choisir le variogramme théorique le plus approprié parmi les modèles théoriques existants, à savoir le modèle linéaire, le modèle exponentiel, le modèle gaussien et le modèle sphérique. Le choix est fait en comparant l'allure de la courbe du modèle théorique avec celle du variogramme empirique estimé à partir des observations. Dans une deuxième étape, à partir de la fonction théorique choisie, les trois paramètres nécessaires du variogramme ; la pépite (*nugget*), la portée (*range*) et le pallier (*sill*) sont estimés par la méthode des moindres carrées ordinaires ou la méthode des moindres carrées pondérées. La section 2.3.1 présente la forme fonctionnelle ainsi que l'allure des différentes fonctions théoriques de variogramme et la méthode d'estimation paramétrique est présentée dans la section 2.3.2.

2.3.1. Modèles théoriques de variogramme

Plusieurs formes de variogramme théorique existent dans la littérature. Les modèles paramétriques de variogramme les plus couramment utilisés sont le modèle linéaire, le modèle exponentielle, le modèle gaussienne et le modèle sphérique. Il existe de nombreux autres modèles, mais ceux présentés ici sont les plus utilisés et sont tous pré-implémentés dans la majorité des logiciels statitiques (*Stata*, *R* ou *SAS*). La Figure II.4 montre les différentes fonctions de variogramme théorique présentées dans cette partie.

Figure II.4: Les différentes modèles de variogramme théoriques



Les paramètres de graphique sont θ_0 : la pépite ou *nugget*, θ_1 : le pallier ou *sill*, θ_2 : la portée ou *range* et *b* : la pente du variogramme linéaire.

Le modèle linéaire

Le modèle le plus simple pour définir le semivariogramme est le modèle linéaire. Ce modèle suppose que le semivariogramme soit une fonction linéaire croissante. Cette fonction linéaire est définie par :

$$\gamma(h;\theta) = \begin{cases} 0, & \|h\| = 0 \\ b\|h\|, & \|h\| \neq 0 \end{cases} \qquad \text{Eq. II.13}$$

b est le taux de croissance constant du variogramme. Le variogramme linéaire n'atteint jamais un pallier, c'est la raison pour laquelle le paramètre b est défini à la place de θ_1. Un semivariogramme linéaire signifie que la variance entre les observations augmente de façon linéaire avec la distance séparant les observations. Comme ce variogramme théorique n'atteint jamais un pallier, il est impossible de définir la distance à partir de laquelle la covariance spatiale entre les observations devient nulle. Mais si cette fonction linéaire de variogramme paraît assez aisée et simple à manipuler, elle explique rarement la réalité.

Le modèle sphérique

La fonction de semivariogramme sphérique est définie par :

$$\gamma(h;\theta) = \begin{cases} 0, & \|h\| = 0 \\ \theta_1\left[1{,}5\left(\frac{\|h\|}{\theta_2}\right) - 0{,}5\left(\frac{\|h\|}{\theta_2}\right)^3\right], & 0 < \|h\| \leq \theta_2 \\ \theta_1, & \|h\| \geq \theta_2 \end{cases} \qquad \text{Eq. II.14}$$

La fonction de covariogramme sphérique est définie comme :

$$C(h;\theta) = \begin{cases} \theta_1, & \|h\| = 0 \\ \theta_1\left\{1 - \left[1{,}5\left(\frac{\|h\|}{\theta_2}\right) - 0{,}5\left(\frac{\|h\|}{\theta_2}\right)^3\right]\right\}, & 0 < \|h\| \leq \theta_2 \\ 0, & \|h\| \geq \theta_2 \end{cases} \qquad \text{Eq. II.15}$$

θ_1 et θ_2 sont le pallier et la portée. Le semivariogramme sphérique atteint le pallier et devient constant à partir de la distance égale à la portée. Il existe donc une portée finie. Le modèle sphérique est particulièrement utilisé pour analyser l'autocorrélation spatiale

lorsque celle-ci décroît presque linéairement avec la distance et devient nulle au-delà d'une distance finie qui est la portée. La portée désigne donc la distance à partir de laquelle la covariance entre les observations disparaît. C'est probablement le modèle de variogramme théorique le plus utilisé dans la pratique.

Le modèle exponentiel

Les fonctions semivariogramme $\left(\gamma(h;\theta)\right)$ et covariogramme $\left(C(h;\theta)\right)$ exponentielles sont définies de la façon suivante :

$$\gamma(h;\theta) = \begin{cases} 0, & \|h\| = 0 \\ \theta_1\left(1 - exp\left(-\frac{\|h\|}{\theta_2}\right)\right), & \|h\| > 0 \end{cases} \qquad \text{Eq. II.16}$$

$$C(h;\theta) = \begin{cases} \theta_1, & \|h\| = 0 \\ \theta_1 exp\left(-\frac{\|h\|}{\theta_2}\right), & \|h\| > 0 \end{cases} \qquad \text{Eq. II.17}$$

θ_1 et θ_2 sont le pallier et la portée de la fonction. Le modèle exponentiel a une allure semblable au modèle sphérique mais le variogramme exponentiel atteint seulement asymptotiquement le pallier. Par conséquent, il n'existe pas de portée finie, mais une portée pratique qui est définie comme la distance à partir de laquelle le variogramme atteint 95% de la variance de l'échantillon. Une autre différence est l'observation de la croissance du semivariogramme exponentiel à un taux décroissant alors que le semivariogramme sphérique croît à un taux constant. Cela signifie que la covariance du modèle exponentiel décroit à un taux moins élevé que la covariance du modèle sphérique. L'utilisation du modèle exponentiel peut être envisagée, dans le cas où la corrélation spatiale disparaît lentement. Par exemple, pour étudier l'autocorrélation des prix immobilier liée à la création d'une nouvelle autoroute. L'existence de l'autoroute a un effet potentiellement positif sur les prix des appartements situés non seulement dans la commune qu'elle dessert mais aussi les prix des appartements situés dans ses communes voisines parce que cette autoroute permet d'améliorer la possibilité de déplacement entre

les communes. Un tel effet persisterait, pour les observations séparées par une longue distance et cet effet dimininue lentement.

Le modèle gaussien

Les équations Eq. II.18 et Eq. II.19 explicitent les fonctions semivariogramme et covariogramme gaussiennes.

$$\gamma(h;\theta) = \begin{cases} 0, & \|h\| = 0 \\ \theta_1\left(1 - exp\left(-\left(\frac{\|h\|}{\theta_2}\right)^2\right)\right), & \|h\| > 0 \end{cases} \qquad \text{Eq. II.18}$$

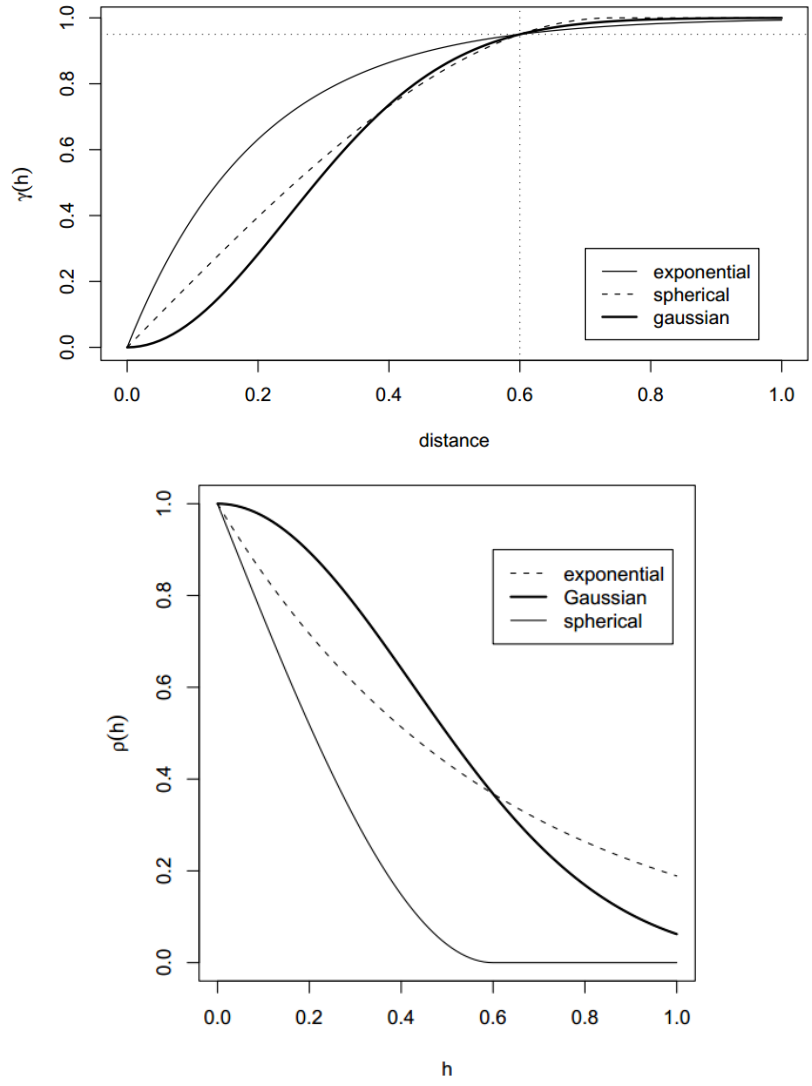
$$C(h;\theta) = \begin{cases} \theta_1, & \|h\| = 0 \\ \theta_1 exp\left(-\left(\frac{\|h\|}{\theta_2}\right)^2\right), & \|h\| > 0 \end{cases} \qquad \text{Eq. II.19}$$

θ_1 et θ_2 sont le pallier et la portée de la fonction. La forme de la fonction gaussienne est semblable à celle du modèle sphérique. De même que le modèle exponentiel, le variogramme gaussien n'atteint qu'asymptotiquement le pallier et on définit la portée par la distance à partir de laquelle le variogramme attient 95% de la variance de l'échantillon. Ce qui différencie le modèle gaussien par rapport au modèle sphérique et au modèle exponentiel est que la fonction semivariogramme gaussienne a la forme d'un « *S* ». Elle présente une forme convexe au départ et devient ensuite concave. Cela signifie que, pour les petites distances, la covariance du modèle gaussien décroît à un taux moins élevé que la covariance du modèle sphérique. Le taux de croissance est faible pour les petites distances, cela signifie que la covariance est très forte parmi le voisinage très proche.

Dans la pratique, le modèle gaussien est utilisé lorsque les données présentent une forte corrélation entre le voisinage proche, cette dépendance étant décroissante et nulle au delà d'une certaine distance. Dans l'exemple de l'étude immobilière, le modèle gaussien peut être utile pour étudier la dépendance spatiale liée à l'existence de stations de métro. Une station de métro a un impact positif sur les prix des biens immobiliers proches de la station. Cette influence décroît avec la distance à parcourir à pieds entre la station et

l'appartement. Si la distance à parcourir à pied va au-delà d'une certaine distance limite, l'individu considère que la station de métro est trop loin pour marcher et n'intègre pas l'existence d'une station de métro proche dans sa détermination du prix de l'appartement.

Figure II.5 : Comparaison le variogramme sphérique, exponentiel et gaussian.



Selon la littérature en finance de l'immobilier, l'analyse de l'autocorrélation spatiale des prix immobiliers utilise souvent la fonction sphérique (Basu et Thibodeau (1998); Hayunga et Pace (2010); Tu, Sun et Yu (2007)) parce qu'elle permet d'obtenir une portée finie, qui est nécessaire pour définir la segmentation de marché.

2.3.2. Estimation paramétrique

A partir du modèle de semivariogramme théorique ($\gamma(h)$) choisi, les paramètres de la fonction semivariogramme empirique ($\hat{\gamma}(h)$) sont estimés par la méthode d'itération (Tu, Sun et Yu (2007)). Ce sont les paramètres qui permettent de minimiser la différence entre ces deux variogrammes à chaque variation de distance $(\gamma(h) - \hat{\gamma}(h))$. Les paramètres $\hat{\theta}$ estimés par moindres carrées ordinaires (MCO) sont définis par :

$$\hat{\theta} = \arg\min_{\theta} \sum_{k=1}^{K} \left[\gamma\big(h(k)\big) - \hat{\gamma}(h(k);\theta)\right]^2 \qquad \text{Eq. II.20}$$

Où θ est le vecteur des trois paramètres à estimer ; la pépite, le pallier et la portée, $\theta = \{\theta_0, \theta_1, \theta_2\}$ et K est le nombre d'intervalles de distance incluses dans la procédure de minimisation.

L'estimation par la méthode des moindres carrées ordinaires est souvent critiquée parce que le nombre d'observations dans chaque intervalle de distance (N_h) est différent. La méthode des moindres carrées pondérées est proposée en pondérant le variogramme de chaque intervalle par le nombre d'observations dans cet intervalle.

2.4. Prévision

Comme pour toutes les méthodes statistiques, il est indispensable d'obtenir une prévision fiable à partir des échantillons collectés. La géostatistique définit le *krigeage* comme une méthode d'interpolation spatiale permettant de prendre en compte la dépendance spatiale et faire de la prévision à partir des données spatiales. Cette méthode se base sur les paramètres estimés du covariogramme et semivariogramme tels que présentés dans la section précédente. La littérature mentionne trois principales méthodes de krigeage dont l'utilisation dépend de l'information statistique de la variable étudiée : le krigeage simple, le krigeage ordinaire et le krigeage universel.

Le krigeage simple est appliqué si le processus spatial de la variable étudiée est stationnaire avec la moyenne statistique connue. Le krigeage simple définit l'écart par

rapport à la moyenne d'une observation $\left(z(s_p) - \mu\right)$ comme la somme des écarts à la moyenne des observations voisines pondérées.

$$z(s_p) = \mu + \sum_{i=1}^{m} w_i \left(z(s_i) - \mu\right) = \sum_{i=1}^{m} w_i \left(z(s_i)\right) + \left(1 - \sum_{i=1}^{m} w_i\right)\mu \qquad \text{Eq. II.21}$$

w_i est la pondération associée à l'observation voisine s_i. Il s'agit alors de déterminer la pondération la plus appropriée. Ce poids, w_i, est déterminé à partir d'un système d'optimisation en minimisant la variance.

Le krigeage ordinaire est destiné à étudier une variable qui suit un processus stationnaire l'espérance est donc inconnue. C'est le krigeage le plus utilisé car il ne nécessite pas de connaître la valeur de l'espérance. Le principe général de l'interpolation spatiale est d'estimer la valeur d'une observation au point s_p, $\left(z(s_p)\right)$, en faisant la moyenne pondérée des valeurs des *m* observations voisines, $\left(z(s_i)\right)$.

$$z(s_p) = \sum_{i=1}^{m} w_i \, z(s_i) \qquad \text{Eq. II.22}$$

Comme dans le krigeage simple, w_i est la pondération associée à chaque observation voisine déterminée à partir du variogramme théorique.

Les deux krigeages présentés ci-dessus sont les plus utilisés. Asli et Marcotte (1995) affirment que ces deux méthodes fournissent des estimations similaires dans les zones avec un nombre important d'observations. Par contre, si le nombre d'observations est faible et sa distribution est étalée dans l'espace, le krigeage simple attribue un poids important à la moyenne globale supposée connue, alors que le krigeage ordinaire attribue le même poids à une moyenne estimée localement. Il existe des krigeages plus avancés autre ceux présentés dans cette partie. Le krigeage universel, comme son nom indique, est la méthode générale destinée à étudier toute variable qui ne suit pas un processus stationnaire. Le krigeage avancé prend en compte l'anisotropie du processus spatial.

Comme la prévision n'est pas l'objectif principal de cette thèse, nous n'avons présenté dans cette partie qu'une brève introduction au krigeage.

3. Géostatistique et finance immobilière

Même si la géostatistique a été développée pour l'étude de gisements miniers, elle est actuellement appliquée dans diverses disciplines comme la météorologie, la chimie, l'urbanisme, l'économie et aussi la finance de l'immobilier. L'étude de marché immobilier physique rencontrant souvent le problème de dépendance spatiale des prix immobiliers, la géostatistique est utilisée pour étudier cette dépendance vis-à-vis des biens voisins. L'étude du degré de la dépendance spatiale permet de mieux estimer la valeur d'un bien immobilier, d'étudier la tendance du marché, de déterminer la segmentation de marché et de choisir les biens immobiliers dans lesquels investir dans le cas d'une diversification de portefeuille immobilier.

Dans la littérature, certains auteurs essayent de comparer la performance de prévision des différentes méthodes utilisées en étude immobilière. Les résultats de Bourassa, Cantoni et Hoesli (2007) et Case, Clapp, Dubin et Rodriguez (2004) montrent que la géostatistique permet d'obtenir de meilleures prévisions que la régression hédonique simple ou que le modèle d'économétrie spatiale. Cependant, la géostatistique est peu utilisée dans les études immobilières, Cela peut être dû aux hypothèses très contraignantes. Diverses questions se posent alors : est-il raisonnable de faire toutes ces hypothèses sur les prix immobiliers ? La géostatistique peut-elle être appliquée à tous les types des données immobilières ? Le fait de choisir le modèle théorique de variogramme est assez abstrait. Il n'existe pas de règle générale de choix. Tout dépend de l'allure du variogramme empirique et de l'objectif de l'étude. On peut alors se demander quel serait le variogramme théorique le plus approprié dans de cas des études immobilières.

Cette partie concerne l'analyse critique de l'application de la méthode géostatistique dans l'étude immobilière, en discutant sur les hypothèses à respecter, en examinant le choix du modèle variogramme théorique, et en regardant la situation pour

laquelle la géostatistique peut fournir des informations intéressantes en étude immobilière.

3.1. Etude géostatistique immobilière

La géostatistique est appliquée principalement avec quatre objectifs : fournir une bonne estimation des prix, améliorer la qualité de la prévision des valeurs de biens immobiliers, déterminer la segmentation de marché et aussi mettre en exergue l'influence d'externalités sur la valeur d'un bien immobilier. La revue de littérature que nous ferons par la suite sera organisée suivant les champs d'application tel que décrits ci-dessus.

Dans la majorité des cas d'étude immobilier, la géostatistique est utilisée pour fournir de meilleures estimations et prévisions de valeurs de biens immobiliers. Si les données immobilières présentent la propriété de dépendance spatiale, la méthode des moindres carrées ordinaires fournit une estimation hédonique biaisée. La géostatistique est donc utilisée afin d'estimer le degré de corrélation spatiale entre les prix immobiliers, ce qui permet d'obtenir une matrice de variance-covariance utilisée par la suite dans la régression par les moindres carrées généralisées afin d'obtenir une estimation non biaisée (Pace, Barry et Sirmans (1998)). Basu et Thibodeau (1998) et Gillen, Thibodeau et Wachter (2001) quant à eux appliquent la géostatistique aux données immobilières afin d'améliorer la qualité de la prévision. Basu et Thibodeau (1998) travaillent sous les hypothèses de stationnarité au second ordre du processus spatial et d'isotropie du variogramme mais Gillen, Thibodeau et Wachter (2001) ignorent l'hypothèse d'isotropie et permettent donc que le semivariogramme puisse varier selon la direction. Chez Basu et Thibodeau (1998), le processus spatial apparaît non stationnaire d'après le semivariogramme de certaines zones, cependant ils ne vont pas plus loin dans leur analyse. Les deux articles utilisent le modèle sphérique pour l'ajustement de ce semivariogramme empirique. Ils expliquent ce choix par le fait que le semivariogramme sphérique contient une portée finie contrairement aux semivariogrammes exponentiel ou gaussien. Par ailleurs, ces deux articles comparent le semivariogramme estimé à partir des prix de transactions avec celui estimé à partir des résidus de l'estimation hédonique, cette comparaison permettant de déterminer la source de la corrélation spatiale. Ils trouvent que la prise en compte des caractéristiques physiques dans la régression hédonique permet de

contrôler la corrélation spatiale dans certaines zones et de réduire la portée estimée sans pour autant supprimer totalement la corrélation. Gelfand, Ecker, Knight et Sirmans (2004) combinent la dimension spatiale et temporelle de façon à prendre en compte la variabilité temporelle.

Autre que l'objectif de l'évaluation de biens immobiliers, la géostatistique est utilisée pour montrer que la segmentation administrative n'est pas un critère suffisant pour la segmentation du marché immobilier. La corrélation spatiale est utilisée comme outil principal pour déterminer la segmentation, et les biens immobiliers sont regroupés dans un même segment si leurs prix sont corrélés. La géostatistique nous permet de définir la distance au delà de laquelle l'autocorrélation spatiale entre les prix immobiliers disparaît, appelée portée, qui permet de segmenter le marché immobilier. La segmentation du marché immobilier à partir du degré de corrélation spatiale est aussi un outil de choix de portefeuille immobilier car il permet d'identifier les biens non corrélés à un portefeuille afin de mieux diversifier celui-ci. Les travaux de Cano-Guervós, Chica-Olmo et Hermoso-Gutiérrez (2003) et de Tu, Sun et Yu (2007) appliquent la géostatistique aux résidus de l'estimation hédonique pour déterminer la segmentation du marché immobilier. Tu, Sun et Yu (2007) travaillent sous les strictes hypothèses de stationnarité et d'isotropie, et utilisent le modèle de semivariogramme sphérique comme modèle théorique d'ajustement. Cano-Guervós, Chica-Olmo et Hermoso-Gutiérrez (2003) ignorent l'hypothèse de stationnarité et ajoutent un modèle de tendance dans le cas où le semivariogramme est non stationnaire. En regardant l'allure de son semivariogramme empirique, le modèle sphérique est choisi comme modèle d'ajustement.

La géostatistique est parfois utilisée pour étudier l'influence d'une nouvelle infrastructure ou l'effet d'une externalité (la pollution, la nuisance sonore, la qualité de l'air, etc.) sur le prix immobilier (Rossi, Mulla, Journel et Eldon (1992)). Fernandez-Aviles, Minguez et Montero (2012) utilisent la géostatistique, avec les hypothèses de stationnarité et d'isotropie, pour mesurer l'influence de la pollution sur le prix résidentiel. Cet article diffère des autres par le fait qu'il ajuste le semivariogramme empirique avec

plusieurs modèles théoriques : l'effet de pépite[3], le modèle exponentiel et le modèle sphérique.

Dans la majorité des cas d'étude immobilière, la géostatistique n'est pas utilisée directement sur les prix de transactions mais plutôt sur les résidus de l'estimation hédonique. Cette méthode considère que la corrélation spatiale est présentée uniquement dans les résidus. Ainsi, la source de corrélation spatiale considérée en géostatistique est la corrélation causée par les variables omises lors de la définition du modèle hédonique. D'autres sources de corrélation, comme la ressemblance des caractéristiques physiques ou le processus de valorisation, ne sont pas prises en compte. C'est une contrainte par rapport au modèle d'économétrie spatiale pour lequel il existe plusieurs modèles d'estimation selon la source de la dépendance spatiale. Dans le cas où la corrélation spatiale est causée par la ressemblance des caractéristiques physiques, le modèle d'économétrie spatiale contient la dépendance des variables exogènes (modèle modèle de variables exogènes décalées - *SLX*). Dans le cas où la corrélation apparaît lors du processus de valorisation des biens, la dépendance spatiale est prise en compte dans la variable endogène (modèle autorégressif spatial - *SAR*). Si la corrélation spatiale est causée par des variables omises lors de la spécification du modèle, la dépendance spatiale est prise en compte dans les résidus (modèle d'erreurs spatiales - *SEM*).

Ainsi, les hypothèses de travail varient suivants les articles. La majorité des auteurs appliquent la géostatistique sous les hypothèses de stationnarité de second ordre du processus spatial et d'isotropie du variogramme. Certains rejettent l'hypothèse d'isotropie et étudient le variogramme directionnel. Certains abordent le problème de processus spatial non stationnaire et d'autres ne vérifient même pas la violation des hypothèses du modèle. Le choix du variogramme théorique est aussi différent. Certains auteurs choisissent le modèle sphérique mais certains préfèrent le modèle exponentiel. Il n'existe pas de règle générale dans le choix ; tout dépend de l'allure du variogramme empirique.

[3] L'effet de pépite pur est le cas où le variogramme fluctue autour d'une valeur constante ce qui signifie qu'il y a indépendance totale des observations.

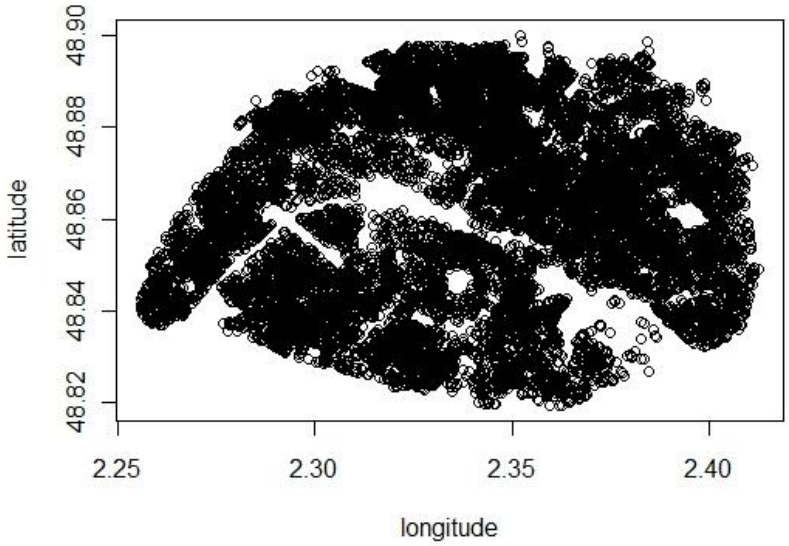
3.2. Processus continu, stationnaire et isotrope ?

La géostatistique est habituellement appliquée sous trois conditions principales: le processus spatial doit être continu et stationnaire et le variogramme isotrope, c'est-à-dire invariant directionnellement. Ces trois hypothèses sont raisonnables pour des données de gisements miniers en géographie mais il semble moins raisonnable de supposer que la distribution spatiale des données immobilières sont stationnaire et que la direction par rapport au centre ville n'a aucun influence sur la détermination de prix de biens immobiliers.

3.2.1. Continuité ou discontinuité

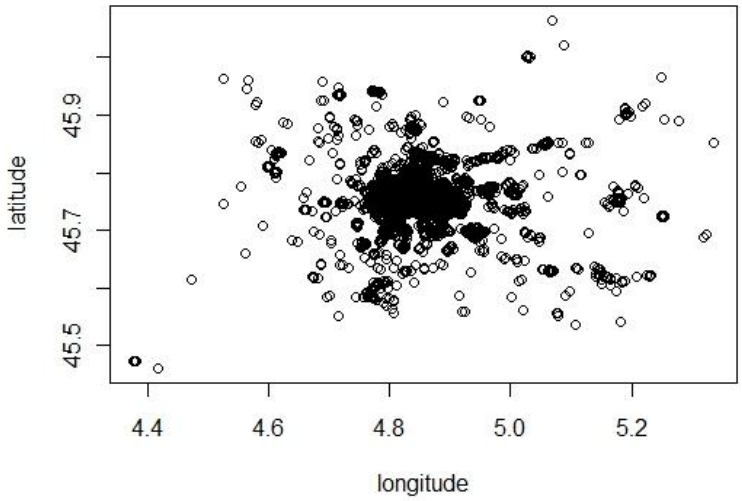
L'hypothèse de la continuité spatiale signifie que les indicateurs de localisation d'observations se distribuent de façon aléatoire dans l'espace continu. Plus précisément, la base de données ne devrait pas être coupée par une frontière ou une barrière. Si le champ d'étude est divisé en plusieurs segments, il faut analyser chaque segment séparément. L'analyse géostatistique, sous l'hypothèse de la continuité, ne pose aucun souci si le nombre de transactions collectées est important et si les observations sont dispersées dans l'espace. Un exemple est celui des données de transactions résidentielles parisiennes en 2007 où il existe environs 28 000 transactions enregistrées et distribuées dans l'espace. Voir le nombre de transactions, nous pouvons supposer que la distribution spatiale des transactions résidentielles parisiennes est continue. En regardant la distribution spatiale des observations de la Figure II.6, la frontière visible est la Seine qui coupe le champ d'étude en deux segments, la partie haute et la partie basse. L'étude géostatistique des biens parisiens doit donc prendre en compte cette discontinuité.

Figure II.6 : La distribution spatiale des transactions résidentielles parisiennes en 2007 (28 828 transactions)

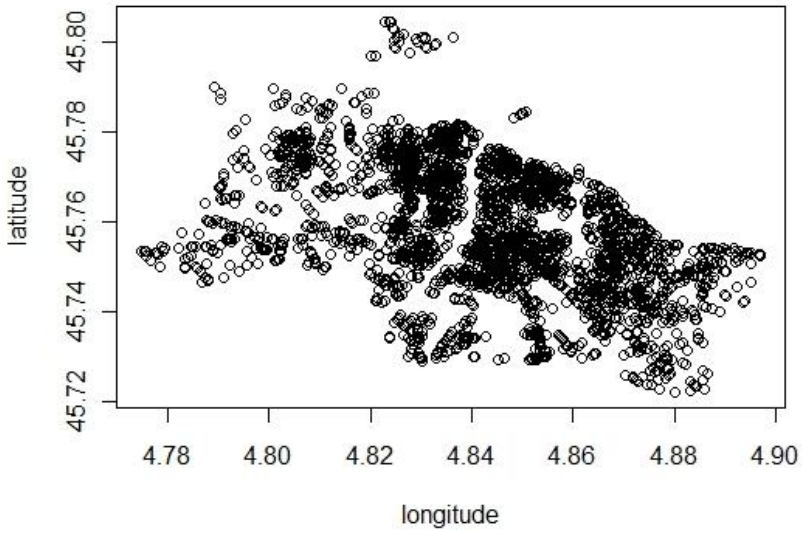


Si on considère d'autre part la base de données des transactions résidentielles à Lyon et ses agglomérations (Figure II.7 – (A)), il est difficile de supposer que les biens résidentiels lyonnais sont distribués de façon continue dans l'espace. Il y a environs 10 000 transactions enregistrées pour l'agglomération de Lyon, la moitié concerne les transactions immobilières à l'intérieur du centre de Lyon (les 9 arrondissements) et l'autre moitié provient des 57 communes restantes. Il est donc irréaliste de supposer que la distribution des transactions immobilières dans l'ensemble de l'agglomération lyonnaise est spatialement continue. En regardant uniquement les transactions du centre de Lyon, l'hypothèse de continuité semble être respectée (Figure II.7– (B)) mais une discontinuité apparait à cause d'une frontière qui divise la base de données en deux segments ; la rive droite et la rive gauche de la Saône. Afin de respecter l'hypothèse de continuité spatiale, l'analyse géostatistique devrait être appliquée à chaque segment.

Figure II.7 : La distribution spatiale des transactions résidentielles dans l'agglomération lyonnaise (A) et Lyon centre (B) en 2007



(A) – **Agglomération Lyon** (10 312 transactions)



(B) – **Lyon Centre** (4 658 transactions)

3.2.2. Stationnarité ou non-stationnarité

La stationnarité spatiale est une deuxième hypothèse souvent imposée en étude géostatistique. Elle est plus contraignante que l'hypothèse de la continuité et elle est peu souvent vérifiée en pratique, surtout dans le cas de la distribution des transactions immobilières. Sous l'hypothèse de stationnarité du second ordre, la moyenne et la

variance de la distribution spatiale sont constantes à travers l'espace et sous l'hypothèse de stationnarité intrinsèque, la moyenne et la variance des variations sont constantes à travers l'espace. Par conséquence, les variogrammes estimés à partir de chaque sous échantillon doivent avoir la même allure. Les biens immobiliers sont très hétérogènes et leurs prix dépendent partiellement de leurs localisations, des externalités et de l'environnement. Gallo (2002) indique que les caractéristiques et les prix des logements diffèrent substantiellement selon leurs localisations. Dans ces conditions, l'estimation d'une relation « globale » entre le prix de logements et ses caractéristiques, où la relation est définie de la même façon sur toute l'aire urbaine étudiée, est susceptible de masquer des différences importantes dans l'espace. Basu et Thibodeau (1998) indiquent que pour certaines zones de son champ d'étude le semivariogramme paraît non stationnaire. Shimizu et Nishimura (2007) étudient la structure du marché immobilier japonais et montrent statistiquement que la structure des prix immobiliers diffère considérablement selon les quartiers.

Le problème de non-stationnarité peut être dû aux différentes raisons. Il paraît possible que les prix de biens immobiliers dans certain segment sont plus corrélés que ceux localisés dans les autres segments. La non-stationnarité peut être dû au fait que les biens immobiliers ne sont pas influencés pas les mêmes externalités. Par exemple, l'influence d'une station de métro sur le prix d'un bien immobilier dépend de la distance entre la station et le bien et le type (résidentiel ou commercial) du bien immobilier (Debrezion, Pels et Rietveld (2007); Grass (1992)). L'existence de la station de métro peut avoir un effet négatif sur le prix des appartements très proche de la station à cause de la nuisance sonore et d'une augmentation du taux de criminalité (Bowes et Ihlanfeldt (2001)). Au contraire, pour les appartements situés un peu plus loin, la station de métro à un effet positif sur les prix des appartements grâce à la facilité de déplacement. Le degré de corrélation spatiale des prix de biens immobiliers situés très proch du métro n'a pas de même niveau que celui des prix de biens situés quelques centaine de mètres plus loin. Deuxième raison, il est peu probable que la moyenne des prix des appartements situés au centre ville soit la même que celle des appartements (de même caractéristiques physiques) situés en dehors du centre-ville. Prenons le cas des transactions immobilières parisiennes. Il est peu réaliste de dire que la moyenne et la variance des prix immobiliers dans le 18ème arrondissement sont les mêmes que celle des biens, de même caractéristiques

physiques, localisés dans le 16ème arrondissement. Tous ces exemples permettent de voir qu'il est irréaliste de faire l'hypothèse de stationnarité spatiale pour la distribution des prix de biens immobiliers.

La stationnarité spatiale peut être obtenue en divisant la base de données en plusieurs segments et en analysant le variogramme par segment. Si le processus spatial est stationnaire, le variogramme obtenu à partir de différents segments devrait avoir la même allure et fournir les mêmes paramètres estimés (la pépite, la portée et le pallier). La discussion de cette hypothèse dans le cadre de notre travail empirique sur la stationnarité spatiale est présentée dans le 0.

3.2.3. Isotropie ou anisotropie

Si le processus spatial est isotrope, la dépendance spatiale entre des biens immobiliers n'est fonction que de la distance les séparant. La pertinence de cette hypothèse dépend de la taille du champ d'étude. Si la base de données concerne des transactions immobilières localisées dans un seul quartier, il est plausible de supposer que la corrélation spatiale ne dépend pas de la direction. Par contre, si la base de données concerne les transactions localisées dans une ville ou une région entière, il faudra au préalable vérifier si l'hypothèse d'isotropie est satisfaite. Plusieurs travaux dans la littérature mentionnent le cas d'anisotropie. Besner (2002) analyse l'évolution des prix de biens résidentiels et montre que la structure de corrélation spatiale ne dépend pas uniquement de la distance séparant les biens immobiliers mais aussi de leur direction. Gillen, Thibodeau et Wachter (2001) analysent l'autocorrélation spatiale des prix de transactions immobilières à Montgomery et montrent que la structure de la dépendance paraît plus importante pour les biens résidentiels dans la direction du centre ville ou dans la direction du quartierdes affaires.

L'hypothèse d'isotropie peut être vérifiée en comparant les variogrammes estimés à partir de plusieurs directions, par exemple, les directions nord-sud et est-ouest. Sous l'hypothèse d'isotropie du variogramme, l'allure de variogramme est invariante directionnellement (Oden et Sokal (1986)).

3.3. Choix de modèle variogramme empirique

Afin de pouvoir analyser la dépendance spatiale et faire un choix d'investissement, il est nécessaire de connaître le niveau de dépendance spatiale pour n'importe quel vecteur de séparation possible (Hayunga et Pace (2010)). Par ailleurs, les méthodes de prévision comme le krigeage ont besoin de variogrammes continus (Tu, Sun et Yu (2007)). Il est donc indispensable d'ajuster un variogramme théorique au variogramme empirique estimé à partir des observations.

Pour choisir le modèle variogramme théorique, l'allure du variogramme empirique est comparé à celui du modèle théorique. Le variogramme empirique doit être régulier et lisse afin de permettre le choix d'un modèle théorique adapté. Dans le cas d'un gisement minier, la présence de minéraux peut être mesurée à n'importe quel endroit, ce qui permet d'obtenir une distribution continue des observations. Par contre, le marché immobilier est peu liquide et le nombre des biens existants sur le marché est limité. Il est donc difficile d'avoir un nombre important de données de transaction de façon à avoir un variogramme empirique suffisamment lisse. Le variogramme obtenu à partir de données immobilières est le plus souvent discontinu et dispersé. Le choix du variogramme théorique dans ce cas est donc plus difficile. Selon Gratton (2002), le choix du variogramme théorique serait plutôt de l'ordre de l'art qu'une science. Ce choix non paramétrique est souvent critiqué par les praticiens de la géostatistique. Certaines études essayent de supprimer cette étape et développent une méthode d'estimation paramétrique sans avoir besoin de connaître préalablement la forme du variogramme théorique (Cherry (1997); Genton et Gorsich (2002)).

3.4. Caractéristiques de localisation inclues dans la régression hédonique

Il est très souvent mentionnée dans l'application géostatistique de ne pas inclure les variables spatiales dans le modèle régression hédonique afin de laisser la partie des résidus prendre toutes les valeurs correspondantes aux caractéristiques de localisation (Dubin (1988); Dubin (1992)). Selon Dubin (1992) il faut omettre toutes les caractéristiques de voisinage et les mesures d'accessibilité lors de la spécification de

modèle hédonique. Un modèle qui intègre la variable spatiale comme une des variables explicatives et analyse à la fois les résidus par la géostatistique peut causer le problème de doublon dans la mesure de la dépendance spatiale. Par contre, l'analyse de variogramme faite par certains auteurs (Basu et Thibodeau (1998); Bourassa, Cantoni et Hoesli (2007); Gillen, Thibodeau et Wachter (2001)) est estimée à partir des résidus de la régression hédonique dans laquelle l'indice de segmentation (un indicateur spatial) est considéré comme une variable explicative. Cette analyse prend la dépendance spatiale en compte deux fois, une fois dans la régression hédonique et une autre fois dans l'analyse du variogramme.

La remarque de Dubin (1992) est discutable si on fait le lien avec les sources de l'autocorrélation spatiale. Comme nous l'avons vu à la section 4.2 du CHAPITRE I, les sources de l'autocorrélation spatiale peuvent être la ressemblance des caractéristiques physiques des biens voisins, le processus d'évaluation du bien ou les variables omises lors de la définition du modèle d'estimation. La dépendance spatiale causée par le processus d'évaluation du bien immobilier est liée à la localisation de ce bien, c'est-à-dire le propriétaire ou l'expert immobilier se renseigne uniquement sur le prix de vente des biens proches de son bien. Ce processus d'évaluation crée un problème de corrélation spatiale uniquement parmi les biens voisins. Si l'on suppose que l'autocorrélation spatiale est causée par le processus d'évaluation du prix des biens immobiliers, l'exclusion des variables spatiales dans le modèle de régression hédonique est nécessaire. Par contre, si l'on considère que la dépendance spatiale est causée par la variable omise lors de la définition du modèle de régression, la prise en compte des variables spatiales dans le modèle d'estimation hédonique ne crée pas de problème de doublon. La variable spatiale prise lors de la définition de la régression hédonique permet de capter la valeur correspondant à la localisation des biens et les résidus présentent le caractère de dépendance spatiale à cause des variables manquantes lors de cette définition. Can (1992) explique l'impact de la localisation sur le prix immobilier en deux niveaux. Le premier correspond aux « effets de quartier », tels que les logements localisés dans le même quartier partageant les caractéristiques du quartier. Les effets de quartier est une influence, au niveau *macro,* sur les prix de tous les biens du quartier. Le deuxième niveau, le niveau *micro*, correspond aux « effets de contiguïté », c'est à dire, l'influence du prix d'un bien sur les prix des autres biens voisins. Ces effets peuvent traverser les

limites des quartiers. En analysant l'idée proposée par Can (1992), l'indicateur de quartier ajouté au modèle de régression hédonique permet de mesurer les effets de quartier et la dépendance spatiale restant dans les résidus de cette estimation hédonique, qui est causée par les effets de contiguïté.

4. Conclusion

Ce chapitre détaille les hypothèses, les éléments nécessaires à l'analyse géostatistique ainsi que leur application dans le cas de l'étude immobilière. Les trois hypothèses, la continuité, la stationnarité et l'isotropie, peuvent être postulés dans le contexte immobilier ; cependant, il faut apporter une attention particulière à la distribution spatiale des observations avant de pouvoir le faire. Sous ces trois conditions, la base de données ne devrait pas être trop large ni trop petite, contiendrait un nombre d'observations raisonnable et sans barrière géographique divisant le champ d'étude en plusieurs segments. Le choix de variogramme théorique est aussi discuté. Si le marché immobilier étudié présente une forte dépendance spatiale parmi les biens proches et si cette dépendance disparaît dès que la distance augmente, le modèle gaussien parait le plus approprié. Si l'on veut intégrer la distance finie à partir de laquelle la corrélation spatiale entre les prix immobiliers disparaît, le modèle sphérique fournissant une portée finie est le plus convenable. Un autre point souvent discuté lors de l'analyse variogramme est la question sur l'exclusion ou non de variables spatiales dans la définition du modèle de régression hédonique. La réponse dépend de la source de corrélation spatiale considérée. Si l'auteur considère que la corrélation spatiale est causée par le processus d'évaluation des biens immobilier, ignorer l'indicateur spatial lors de la définition hédonique est plausible. Par contre, si l'on estime que la variable omise est une cause de la dépendance spatiale, l'inclusion de l'indicateur spatial permet de contrôler certaines causes de la corrélation spatiale.

L'avantage de la géostatistique est qu'elle permet de déterminer directement le niveau de corrélation spatiale en fonction de la distance sans avoir besoin de définir ex-ante la relation entre le voisinage et les éléments de la matrice de pondérations spatiales. Par contre, elle est plus difficile à mettre en œuvre et demande plus de temps de calcul

que l'économétrie spatiale. Cette méthode nécessite une connaissance en géographie et nécessite de disposer des coordonnées spatiales de chaque observation. Par conséquent, l'analyse géostatistique des données immobilière doit être faite avec beaucoup d'attention. Cette méthode a été initialement développée pour l'analyse de données minières dont les caractéristiques sont très différentes de celles des données immobilières. La géostatistique permet d'analyser, soit une corrélation spatiale des prix immobiliers causée par le processus d'évaluation, soit une corrélation spatiale des résidus de la régression hédonique causée par les variables omises lors de la définition de modèle hédonique. Une autre source de la corrélation spatiale qui est la ressemblance des caractéristiques physiques de biens voisins, ne peut être modélisée par la géostatistique mais par d'autres modèles plus adaptés d'économétrie spatiale.

CHAPITRE III SPATIAL AND TEMPORAL NON-STATIONARY SEMIVARIOGRAM ANALYSIS USING REAL ESTATE TRANSACTION DATA

Abstract: The geostatistic spatial statistical methodology is used to correct spatial autocorrelation problems. This model is developed under two common assumptions made to allow global homogeneity: spatial continuity and spatial stationarity. Different research fields such as geography, environmental science, and computer science usually account for violation of the second assumption (spatial stationarity), but no article works under non-stationary conditions in real estate research. This article is probably the first attempt to examine the violation of the stationary assumption in space and time using transaction prices from 1998 to 2007 of Parisian residential properties situated 3 kilometers around the *Arc de Triomphe* and the *Place d'Italie*. By comparing 1-year to 10-year experimental semivariograms, we find variability in spatial structure over time. Likewise, a violation of the spatial stationarity assumption is examined by comparing semivariograms obtained from each 90-degree rotating data segment. Our results show that we should not compute a single common variogram for the region of interest. An adequate sample size should be applied to ensure local homogeneity and local stationarity. Moreover, we attempt to identify the cause of semivariogram variability, and our results show that two levels of spatial correlation effects should be distinguished: "neighborhood effects" that may partly be captured by including arrondissement or district number as an explanatory variable in hedonic regression and "adjacency effects" that cause spatial correlation among the residuals.

Keywords: non-spatial stationarity, spatial autocorrelation, geostatistic, hedonic index

1. Introduction

Because real estate prices present a spatial dependence structure, property value depends partly on the value of its neighbors and the hedonic regression model customarily used to estimate housing prices does not take into account this spatial dependence structure. Geostatistics may be considered as an extension method that provides statistic tools for spatial data analysis. This study supposes that real estate values are divided in two parts. The first part contains a value of physical characteristics and the second part corresponds to a value of spatial characteristics. Considering that property shares the same influences from location factors with its neighbors, only spatial characteristics are considered as a source of spatial interdependence among property prices. Therefore, traditional hedonic regression is used to determine the implicit values of physical characteristics. Then, in the second step, the geostatistic method is applied to the residuals of hedonic regression to analyze the spatial dependence caused by spatial characteristics.

An important point to be underlined is that geostatistic methodology is developed under two main assumptions: spatial continuity and spatial stationarity. The first assumption is assumed to estimate a continuity variogram and to provide a reliable fitted variogram model. The second one, and the main point of our study, is spatial stationarity. Under this assumption, a semivariogram or a covariogram obtained from any data segments should have the same shape and the same estimated parameters. Because real estate data is heterogeneous and spatial factors affecting property prices may vary by location, assuming spatial stationarity for real estate data seems difficult. The objective of this paper is to examine the violation of the stationary assumption in space and time. This non-stationary problem is mentioned in previous studies of several research fields except real estate research. This article is probably a first attempt to point out the violation of this stationary assumption. Ignorance of the non-stationary problem may lead to a bias estimated property price, an under or overestimated range, and faulty price predictions. Moreover, after experimental semivariograms obtained from each segment are shown to be different, the sensibility range analysis is applied to specify the factor that may cause this spatial variability. Finally, we distinguish two levels of spatial correlation effects:

"neighborhood effects" that may be partly captured by including arrondissement or district number as an explanatory variable in hedonic regression and "adjacency effects" that cause spatial correlation among the residuals.

This study is divided into three steps. For the first step, hedonic regression is applied with whole data to estimate physical characteristic values.

In the second step, a variability of spatial structure over time is examined using transaction prices of residential data from the Paris notary database from 1998 to 2007. We compare the 1-year semivariogram obtained from annual data to the 10-year global semivariogram estimated from whole data. Under this stationary assumption, all 1-year semivariogram should have a similar shape with a similar estimated range. Our results cannot confirm whether the spatial structure of our data is stationary over time because we observe the same shape of a semivariogram but the estimated ranges are different. This variability in the estimated range may be the result of the development of Paris's real estate market. A longer estimated range means higher price correlation. From 1998 to 2007, the Paris real estate market developed and prices became more correlated.

The third step examines the spatial stationarity assumption. To reduce calculation time, we take only the properties located 3 kilometers around two main points of Paris: the *Arc de Triomphe* and the *Place d'Italie*. The data around each main point are divided in 36 rotating segments and then semivariograms obtained from each segment are compared. Under the spatial stationarity assumption, all 36 semivariograms should have the same shape and the same estimated range should be obtained. The results confirm the notion that real estate data are heterogeneous, and spatial stationarity could not be assumed because our 36 semivariograms are dissimilar. Semivariograms depend on market structure, environmental factors, and externalities. We remark that local spatial stationarity is assumed for each data segment to allow the local application of geostatistic method.

To complete our study, after we point out the problem of non-stationary of spatial process, spatial stationarity could not be assumed for all transaction prices of Parisian apartments. The logarithmic transformation proposed by Kerry and Oliver (2007a) that reduces heterogeneously the semivariogram shape is also applied in our study. Despite

several data transformation attempts (square, square root, logarithmic), we could not obtain completely homogenous semivariograms. Therefore, we attempted to identify a factor that may cause this spatial variability by including other characteristics as explanatory variables in the hedonic regression, such as age of seller and buyer, *arrondissement* dummy, and district dummy. We find that the dummy for property location reduced the semivariogram variability and increased the similarity of estimated ranges.

This article is structured as follows. The next section discusses some literature reviews on non-stationary spatial processes. The papers in other research fields such as geosciences, geography, and urban studies that mentioned the violation of spatial stationarity are presented. Then, the third section describes our database. The fourth section concerns the methodology of this paper, and describes the traditional hedonic regression used to evaluate physical characteristics and the geostatistic model used to study the violation of the stationary assumption in space and time. Our empirical results are presented and discussed in the fifth section. The sixth section corresponds to a semivariogram range sensibility analysis and the final section concludes the paper.

2. Literature review

In several papers with different objectives, a geostatistic approach is applied using real estate data. Basu and Thibodeau (1998) used a geostatistic method to illustrate spatial autocorrelation among transaction prices within a submarket. Tu, Sun and Yu (2007) used a geostatistic model to define market segmentation to replace the ad hoc administrative boundary. Gillen, Thibodeau and Wachter (2001) investigated the validity of the anisotropy assumption. To apply a geostatistic model, two important assumptions are frequently used: spatial continuity and spatial stationarity (Cressie (1991)). Spatial continuity supposes that spatial data are distributed continuously. This first assumption is plausible for a dataset with a large number of observations. The second assumption, spatial stationarity, supposes that the mean and variance of each residual distribution should be constants across all regions of interest. Therefore, that this stationary

assumption could hold for heterogeneous real estate data is difficult to believe. Cressie (1991) noted that, assuming that spatial locations of data occur regularly, as they do in a time series model, is simply not reasonable. The violation of the spatial stationarity assumption is less common in real estate research. We remark that all previous works on real estate always utilized the spatial stationarity assumption (Basu and Thibodeau (1998); Bourassa, Cantoni and Hoesli (2007); Can (1990); Gillen, Thibodeau and Wachter (2001); Tu, Sun and Yu (2007)).

However, many articles in geography, geosciences, or the statistic field analyze non-stationary variogram situations (Atkinson and Lloyd (2007); Corstanje, Grunwald and Lark (2008); Ekström (2008); Ekström and Luna (2004); Fernández-Casal, González-Manteiga and Febrero-Bande (2003); Haslett (1997); Kerry and Oliver (2007a); Strebelle and Zhang (2005)). Atkinson and Lloyd (2007) and Corstanje, Grunwald and Lark (2008) split their dataset into several segments and showed that semivariogram could vary locally. Atkinson and Lloyd (2007) showed large different results between the global semivariogram obtained from whole data analysis and the local semivariogram obtained from subregional analysis. Corstanje, Grunwald and Lark (2008) confirmed the heterogeneity of variograms and illustrated the different shapes of local variograms obtained from different data segments. Ekström and Luna (2004) noted that the assumption of stationary is easily rejected through an empirical application. They proposed the subsampling method, which is a non-parametric way to estimate the variance based on non-stationary spatial data. Kerry and Oliver (2007a) and Kerry and Oliver (2007b) argued that asymmetry in the distribution of data has an effect on the form of the variogram and proposed data transformation to address the variation in the shape of the variogram. Two transformation functions are proposed to reduce asymmetry: logarithmic and square root. They also noted that the transformation to square root is generally regarded as suitable whereas skewness of data is modest (low asymmetry) and the logarithmic is used for larger departures from a symmetric distribution (high asymmetry). Leuangthong and Deutsch (2005) presented a more complicated model to work under a non-stationary variogram using the Bayesian formulation.

3. Data

This paper studies spatial correlation using transaction prices for Parisian residential property from the *Paris Notaire Service* (PNS) database. From 1998 to 2007, the notary data contained 367,035 observations with over 100 different characteristics. Despite a large number of characteristics, we cannot take into account all of them in our regression. Only certain essential characteristics are considered, including surface, apartment type, number of rooms and bathrooms, floors, apartment equipment (elevator, terrace, garden, parking, basement and extra room[4]), construction period, transaction date, and spatial characteristics (street type, arrondissement, and Cartesian coordinator) of the property under consideration. Table III.1 provides a summary of all retained characteristics for our study.

To exclude irregular transactions, general data cleaning proceeded as follows. We keep only transactions with prices starting from €20,000 to €3,000,000 and for apartment areas between 8 m^2 and 250 m^2, and exclude transactions with prices per square meter outside the interval from €1,000 to €20,000. Only transactions in case of "*vente de gré à gré sans viager en plein propriété*" are taken into account in our sample, i.e., only transaction between two private householders (*vente de gré à gré*); therefore, negotiations with commercial householders or by auction are not considered. Likewise, sales against life annuities are excluded (*sans viager*). Properties should be sold under unrestricted conditions (*plein propriété*), i.e., properties sold under credit bail or under construction are excluded.

[4] Called "*Chambre de service*" or "*Chambre de bonne*" in French, it is a type of French apartment corresponding to a single room found on the top floor of an apartment building and is only accessible using a staircase. Initially, these rooms were intended as bedrooms for domestic workers such as maids.

Table III.1: Descriptive statistics, in-sample 325,531 Parisian residential transaction price from 1998 to 2007.

Variables	Description	Number	Percentage	Cumulative Percentage
Period (Period of construction)				
Bf1850	Before 1850	15,754	4.84	4.84
1850_1913	1850-1913	124,425	38.22	43.06
1914_1947	1914-1947	51,677	15.87	58.93
1948_1969	1948-1969	44,020	13.52	72.45
1970_1980	1970-1980	37,255	11.44	83.89
1981_2000	1981-2000	16,695	5.13	89.02
Af2001	After 2001	3,570	1.1	90.12
NA		32,135	9.87	100
Parking (Existence of parking)				
0	Without parking	273,465	84.01	84.01
1	With parking	52,066	15.99	100
ExtraRoom (Existence of extra room)				
0	Without extra room	313,931	96.44	96.44
1	With extra room	11,600	3.56	100
AprtTyp (Apartment Type)				
AprtStd	Standard apartment	316,042	97.24	97.24
AprtDu	Duplex apartment	8,979	2.76	100
L_Elev (Floor / Existence of elevator)				
L0	Ground floor	28,223	8.67	8.67
L123	1st, 2nd, or 3rd floor	157,533	48.39	57.06
L456naElev	4th, 5th, or 6th floor with unrecognized elevator	71,914	22.09	79.15
L456w/oElev	4th, 5th, or 6th floor without elevator	11,972	3.68	82.83
L456wElev	4th, 5th, or 6th floor with elevator	30,459	9.36	92.19
L7plusnaElev	7th floor and higher with unrecognized elevator	6,842	2.1	94.29
L7plusw/oElev	7th floor and higher without elevator	284	0.09	94.38
L7pluswElev	7th floor and higher with elevator	18,304	5.62	100
NbRoom_NbBath (Number of rooms / Number of bathrooms)				
p1s0	1 room without bathroom	9,037	3.02	3.02
p1s1	1 room with 1 bathroom	62,573	20.91	23.93
p1s2	1 room with 2 bathrooms	24	0.01	23.94
p2s0	2 rooms without bathroom	17,380	5.81	29.75
p2s1	2 rooms with 1 bathroom	88,121	29.45	59.20
p2s2	2 rooms with 2 bathrooms	292	0.1	59.29
p3s0	3 rooms without bathroom	7,684	2.57	61.86

Variables	Description	Number	Percentage	Cumulative Percentage
p3s1	3 rooms with 1 bathroom	56,433	18.86	80.72
p3s2	3 rooms with 2 bathrooms	2,084	0.7	81.42
p4s0	4 rooms without bathroom	1,982	0.66	82.08
p4s1	4 rooms with 1 bathroom	24,649	8.24	90.32
p4s2	4 rooms with 2 bathrooms	6,964	2.33	92.65
p5s0	5 rooms without bathroom	1,085	0.36	93.01
p5s1	5 rooms with 1 bathroom	11,229	3.75	96.76
p5s2	5 rooms with 2 bathrooms	9,693	3.24	100
Terrace (Existence of Terrace)				
0	Without terrace	316,643	97.27	97.27
1	With terrace	8,888	2.73	100
Garden (Existence of garden)				
0	Without garden	323,203	99.28	99.28
1	With garden	2,328	0.72	100
Basement (Existence of basement)				
0	Without basement	84,592	25.99	25.99
1	With basement	240,939	74.01	100
Semester				
98s1	S1 of 1998	13,108	4.03	4.03
98s2	S2 of 1998	15,155	4.66	8.69
99s1	S1 of 1999	16,702	5.13	13.82
99s2	S2 of 1999	17,883	5.49	19.31
00s1	S1 of 2000	16,418	5.04	24.35
00s2	S2 of 2000	15,641	4.8	29.15
01s1	S1 of 2001	15,399	4.73	33.88
01s2	S2 of 2001	15,300	4.7	38.58
02s1	S1 of 2002	15,357	4.72	43.30
02s2	S2 of 2002	15,147	4.65	47.95
03s1	S1 of 2003	15,698	4.82	52.77
03s2	S2 of 2003	16,598	5.1	57.87
04s1	S1 of 2004	17,149	5.27	63.14
04s2	S2 of 2004	17,941	5.51	68.65
05s1	S1 of 2005	17,964	5.52	74.17
05s2	S2 of 2005	17,866	5.49	79.66
06s1	S1 of 2006	17,062	5.24	84.90
06s2	S2 of 2006	16,577	5.09	89.99
07s1	S1 of 2007	16,681	5.12	95.11
07s2	S2 of 2007	15,885	4.88	100
Total		**325,531**		

Number of observations and percentage of each dummy variable in a hedonic regression model, dummy variable in gray is used as a reference.

Furthermore, because certain characteristics contained too many categories, further data preparation includes characteristic clustering. The number of rooms starts from 0 to 14. To reduce the number of dummy variables used by the regression, we cluster this characteristic in only five groups (1 to 5plus). The number of bathrooms contains a value from 0 to 9. Apartments with more than 2 bathrooms represent only 0.61%; therefore, the number of bathrooms is grouped into only three categories (0, 1, and 2plus). There are 45 floor categories in our data. Of course, we cannot include 45 dummies in the regression, and even if doing so was possible, floor cannot be considered a continuous value. Two reasons justify this decision. First, the relationship between price and floor is not linear. Normally, we find an important price difference between a ground floor apartment and a 1st floor apartment; however, this difference is smaller between a 3rd floor apartment and a 4th floor apartment. Second, it is not evident that an apartment situated on a higher floor has a higher price if there is no elevator, indicating that floor and existence of elevator may be correlated. To avoid a correlated regressor problem, a crossed variable for floor/existence of elevator is created, and details are provided in the section on hedonic regression. Therefore, we cluster 45 floor categories into four groups; *F0* (ground floor), *F123* (1st, 2nd, and 3rd floors), *F456* (4th, 5th, and 6th floors) and *F7plus* (7th floor and higher). These four groups help reduce the number of variables in the regression and are used to create a crossed variable.

Missing data is another important problem. Two methodologies related to this problem are proposed in previous studies: exclusion or imputation. Concerning the principal variables influencing price determination (surface, number of rooms, number of bathrooms, and floor), unrecognized categories are excluded. Regarding apartment type and equipment, the imputation is applied to replace some missing data. An unrecognized apartment type is replaced by a standard apartment. For apartment equipment, we observe that apartments in Paris seldom have parking, gardens, terraces, extra rooms, or basements. Thus, the unrecognized categories are replaced by the inexistence of such equipment. The two characteristics construction period and existence of an elevator present a high percentage of unrecognized cases, and we decided to keep these samples and define missing data as the unrecognized category.

After the data cleanup procedure, our sample is reduced to 325,531, the average price is €224,855, average surface area is 52 m^2, and average price per square meter is €4,023. Table III.2 provides a summary of our data and the annual data, and Table III.3 shows a summary of the data by year. Our sample has approximately 30,000 data points for each year, and is adequate in providing reliable annual regression results. From 1998 to 2007, we observe that transaction price and price per square meter increase, as does standard deviation with price levels, indicating strong development of the Parisian apartment market since 1998 and higher price fluctuations. From 2003 to 2005, the average price of Parisian apartments increased by 15%, and average price per square meter confirms this argument, which increased from €2,824 in 1998 to €3,879 in 2003 and then doubled to reach €6,194 in 2007.

Table III.2: Data Summary

Variable	Number of Observations	Mean	Std. Dev.	Min	Max
Price	325,531	224,855.2	225,826.6	20,000	3,000,000
Surface	325,531	52.5	34.2	8	250
price_m2	325,531	4,023.4	1,874.5	1,000	20,000

Regarding data distribution in each arrondissement (Table III.4), the majority of our apartments in our sample are located in the 15th, 16th, 17th, and 18th arrondissements. These four arrondissements are well known as residential districts of Paris. A few apartments located in the 1st to the 4th arrondissements because these four arrondissements contain many offices and administrative buildings.

Table III.3: Data Summary by Year

Price	Year	Number Observations	Mean	Annual variation of mean	Std. Dev.
	1998	28,263	140,520.9		132,994.0
	1999	34,585	147,183.6	4.74%	143,086.8
	2000	32,059	165,112.1	12.18%	170,728.5
	2001	30,699	173,461.8	5.06%	176,550.3
	2002	30,504	186,855.4	7.72%	183,731.1
	2003	32,296	216,679.5	15.96%	203,070.2
	2004	35,090	249,496.6	15.15%	226,697.2
	2005	35,830	287,912.8	15.40%	254,005.5
	2006	33,639	313,950.3	9.04%	273,734.8
	2007	32,566	343,535.6	9.42%	301,704.3
Price / m2	**Year**	**Number of Observations**	**Mean**	**Annual variation of mean**	**Std. Dev.**
	1998	28,263	2,428.4		972.6
	1999	34,585	2,631.1	8.35%	1,087.2
	2000	32,059	2,915.6	10.81%	1,219.3
	2001	30,699	3,148.6	7.99%	1,295.8
	2002	30,504	3,398.9	7.95%	1,309.1
	2003	32,296	3,879.1	14.13%	1,390.3
	2004	35,090	4,419.1	13.92%	1,456.0
	2005	35,830	5,122.3	15.91%	1,555.4
	2006	33,639	5,668.8	10.67%	1,624.8
	2007	32,566	6,194.5	9.27%	1,793.5
Surface	**Year**	**Number of Observations**	**Mean**	**-**	**Std. Dev.**
	1998	28,263	54.1		33.6
	1999	34,585	52.3		32.8
	2000	32,059	51.7		33.2
	2001	30,699	50.6		32.0
	2002	30,504	50.8		32.4
	2003	32,296	52.3		34.0
	2004	35,090	53.3		35.3
	2005	35,830	53.6		36.0
	2006	33,639	53.0		36.1
	2007	32,566	53.1		35.6

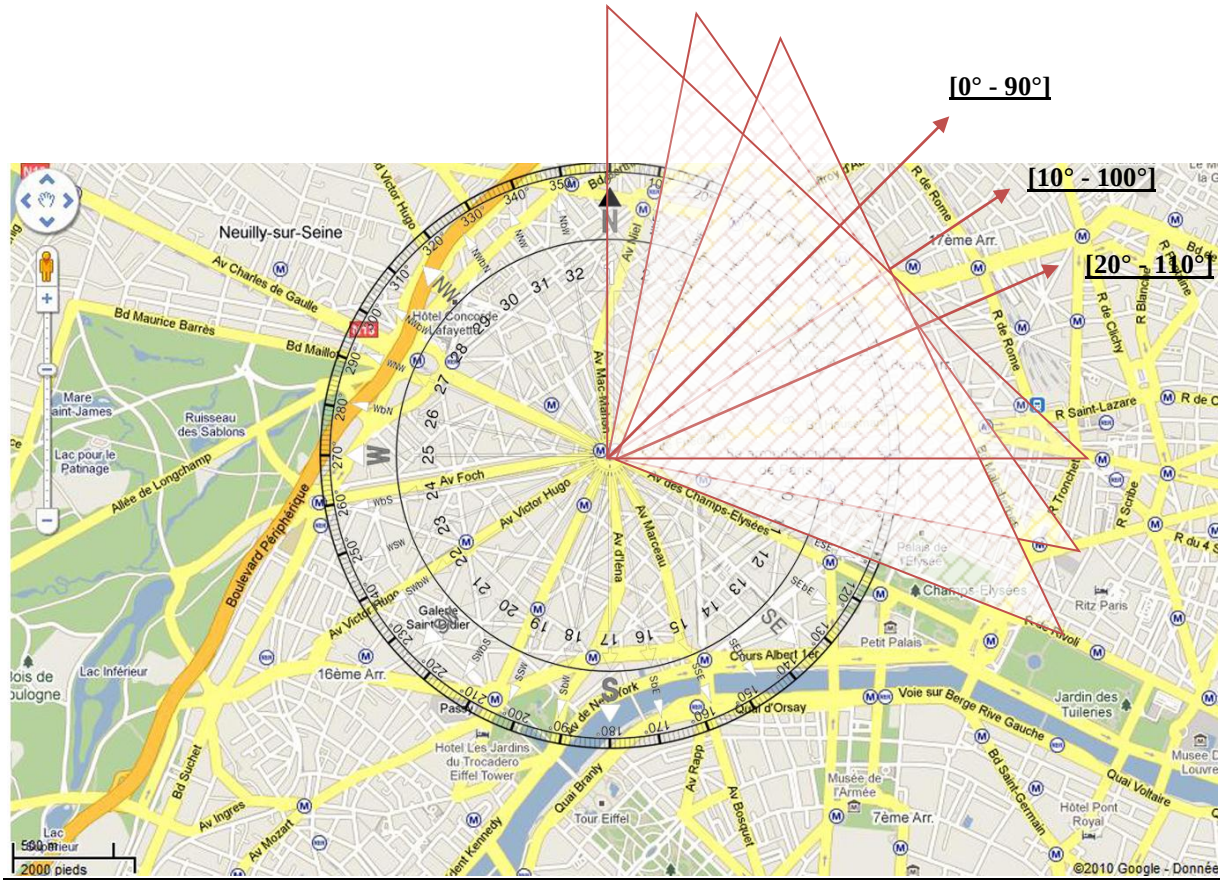
Table III.4: Number of observations by arrondissement

Arron	Freq.	Percent	Cum.
0	223	0.07	0.07
1	2,794	0.86	0.93
2	4,841	1.49	2.41
3	7,437	2.28	4.7
4	4,910	1.51	6.21
5	8,022	2.46	8.67
6	6,581	2.02	10.69
7	8,683	2.67	13.36
8	7,254	2.23	15.59
9	11,735	3.6	19.19
10	16,428	5.05	24.24
11	28,192	8.66	32.9
12	18,289	5.62	38.52
13	17,123	5.26	43.78
14	17,113	5.26	49.04
15	33,562	10.31	59.35
16	26,395	8.11	67.45
17	28,118	8.64	76.09
18	34,694	10.66	86.75
19	20,552	6.31	93.06
20	22,585	6.94	100
Total	325,531	100	

Our study proceeded in three steps. In the first step, a hedonic regression is applied to the entire dataset to estimate physical characteristic values. In the second step, a geostatistic method is used to examine a violation of *time stationary* assumptions. We compare a 10-year semivariogram with an annual one. To reduce the semivariogram computation time, our data are reduced. Only two mark points of Paris are selected: the *Arc de Triomphe* and the *Place d'Italie*. Properties located 3 kilometers around these two mark points are included in our new database (Figure III.2). We consider that 3 kilometers is far enough for analyzing the correlation problem, i.e., the prices of two properties separated by more than 3 kilometers should not be correlated. In the third step, to examine a violation of the *spatial stationarity* assumption, we compare the semivariogram results obtained from different data segments. Therefore, the sample is split into 36 rotating equal segments. Each window is 90 degrees wide and the compass is

rotated 10 degrees from one window to the next one. Figure III.1 provides a better idea of the data segmentation around the *Arc de Triomphe*, which is set as the center point. The first data segment is 0°–90°, i.e., properties located in an area from 0° to 90° of the compass are included in this segment. The second segment is rotated 10° from the first segment, in a clockwise direction and is called 10°–100°. A new segment is always rotated 10° from a previous one. This procedure is applied completely around the *Arc de Triomphe*, and our last segment is 350°–80°. A semivariogram is estimated for all 36 rotating segments and this segmentation concept is applied to the second mark point, *Place d'Italie.*

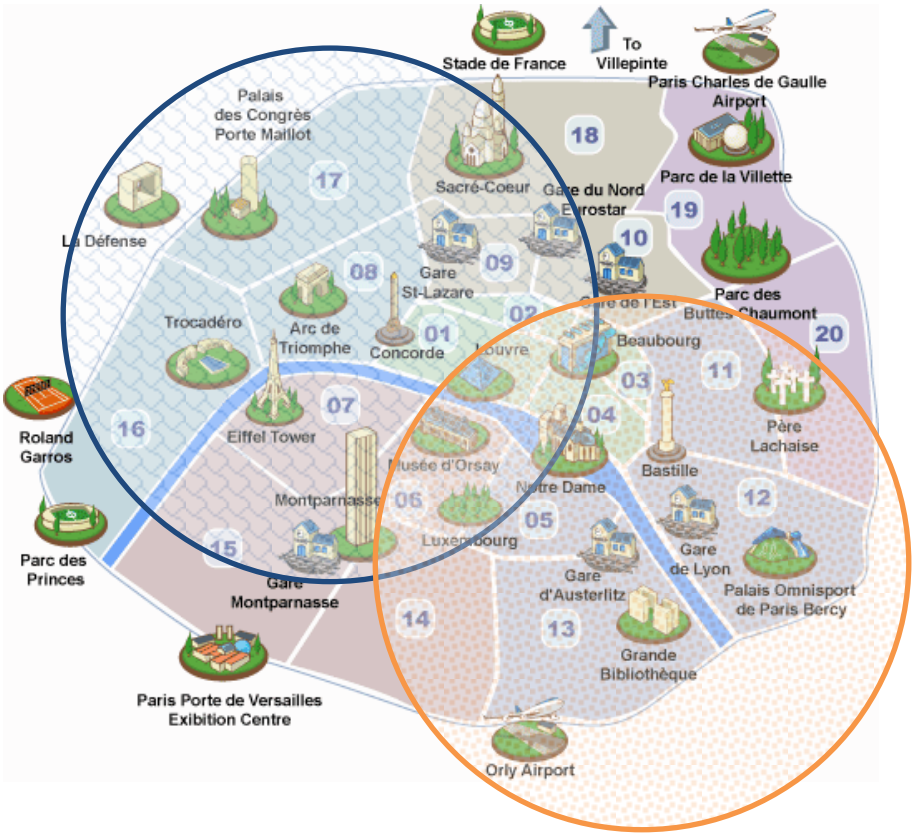
Figure III.1: Data segmentations



The *Arc de Triomphe* and the *Place d'Italie* are selected as two mark points for this study, for several reasons. First, the *Arc de Triomphe* is a very large roundabout in Paris, and 12 avenues depart from this monument. The famous *Avenue des Champs Elysées* is one of these 12 avenues that link the *Arc de Triomphe* and the *Place de la Concorde*. Our data segmentation based on the direction from the *Arc de Triomphe* allows an analysis of property price correlations along each avenue. We may find high price correlation for certain avenues and price differences for another. Second, our database of properties is situated 3 kilometers around the *Arc de Triomphe* and includes properties in the 1st, 8th, 9th, 15th, 16th, 17th, and 18th arrondissements. This study also confirms the previous idea of price correlation among properties in different administrative boundaries; therefore, price prediction based on these pre-defined boundaries is not the most reliable (Tu, Sun and Yu (2007)). Figure III.2 shows a map of 20 arrondissements in Paris and of the location of the *Arc de Triomphe*. The third reason is that there are many types of properties 3 kilometers around the *Arc de Triomphe*. Hence, several residential categories are present in our data; luxury residents are situated near the *Avenue de Champs Elysées*, high-end residences are around the *Parc de Monceau*, working class residents live in the 16th arrondissement, and middle-class residents live north of the 17th arrondissement. These various types of properties included in our data help illustrate the different price structures in various regions of interest.

Place d'Italie, our second mark point, is located in the southeast area of Paris in the 13th arrondissement. The dataset of locations 3 kilometers around the *Place d'Italie* takes into account the properties located in the 4th, 5th, 6th, 12th, 13th, and 14th arrondissements. This data set regroups a variety of property types. Condominiums located in 13th arrondissement, luxury apartments in Haussemanian buildings are located in the 4th, 5th, and 6th arrondissements, and standard apartments are located in the 12th and 14th arrondissements.

Figure III.2: Paris map and selected area of study



4. Methodology

Our research proceeds in three steps. First, we estimate the value of apartment characteristics using hedonic regression. The second and third steps intend to examine a violation of time stationary and spatial stationarity hypothesis. Our data are divided into 36 rotating segments relative to the *Arc de Triomphe* and the *Place d'Italie*. The semivariogram is calculated for each data segment. Semivariogram shapes and estimated ranges are then compared. Under a stationary assumption, we should have the same semivariogram structure and the same estimated range for all data segments.

4.1. Hedonic regression

One of the dominant real estate valuation methods is hedonic regression. Rosen (1974) argued that property is purchased as a tied package of characteristics and property price is valued according to the utility of these characteristics. The hedonic regression result provides an implicit price of the characteristics package associated with property. We model a relationship between apartment price and apartment characteristics that can be defined as a linear function (Eq. III.1).

$$P = \beta_0 + \beta X + \varepsilon \qquad \text{Eq. III.1}$$

where P defines a vector of transaction prices per square meter, β_0 denotes a constant indicating a value of a reference apartment, X is a characteristics matrix, β defines a vector of estimated hedonic coefficients, and ε is a residuals vector. Some previous studies applied an alternative hedonic function such as log, semi-log, and Cox-Box transformations. Because the results obtained in our study from linear regression and log transformation have the same magnitude; we decided to use the linear regression result.

For statistical and economic reasons, price per square meter (P) is selected as a dependent variable. On statistical grounds, dividing transaction price by surface helps correct the size effect because prices tend to be higher as surface area increases. On economic grounds, using price per square meter as a dependent variable permits the capture of a special type of Parisian apartment. In particular, the price per square meter of a one-room apartment with 1 bathroom (called a studio) is always higher than the price per square meter of other categories. The results obtained by our hedonic regression confirm this situation. In addition, our background experience shows that investors are more interested in price per square meter than transacted price.

Each year, we run two hedonic regressions. The first hedonic regression does not take into account any spatial characteristic (arrondissement and road type). We suppose that omitting (missing) variables may cause spatial correlation in the errors terms and a geostatistic method is applied to the error terms of hedonic regression. For the second hedonic model, presented in the section 6 (*Semivariogram range sensitivity analysis*), we follow the idea of Can (1992), who distinguished the "neighborhood effects" and the

"adjacency effects." An arrondissement index is included in the hedonic regression to isolate the value of shared neighborhood characteristics on apartment price and an error term represents price correlation caused by the spatial spill-over effect. Additional details are provided in section 6.

Physical characteristics matrix (*X*) is composed of 10 variables, defined as follows.

- *NbRoom* (Number of rooms), measured using five dummies: *NbRoom1*, *NbRoom2*, *NbRoom3*, *NbRoom4*, and *NbRoom5plus*, respectively, for 1, 2, 3, 4, or 5 or more rooms. A crossed variable is created for this characteristic.
- *NbBath* (Number of bathrooms), measured by three dummies: *NbBath1*, *NbBath2*, and *NbBath3plus*, respectively for 1, 2, or 3 or more bathrooms. A crossed variable is created for this characteristic.
- *Floor*, measured by four dummies: *Floor0* (apartment situated on the ground floor), *Floor123* (apartment situated on the 1st, 2nd or 3rd floor), *Floor456* (apartment situated on the 4th, 5th or 6th floor), and *Floor7plus* (apartment situated on the 7th floor or higher). A crossed variable is created for this characteristic.
- *Elevator* (Existence of elevator), measured by three dummies: *Elev_Y* (Yes), *Elev_N* (No), and *Elev_NA* (unrecognized). A crossed variable is created for this characteristic.
- *AprtTyp* (Apartment Type), measured by two dummies: *AprtStd* (Standard Apartment) and *AprtDu* (Duplex[5] Apartment). *AprtStd* is removed as a reference.
- *Period* (Period of construction), measured by eight dummies: *Na* (for unrecognized period of construction), *Bf1850* (Before 1850), *1850_1913*, *1914_1947*, *1948_1969*, *1970_1980*, *1981_2000*, and *Af2001* (after 2001). *1948_1969* is removed as a reference.
- *Parking* (Existence of parking), a dummy to control whether an apartment has parking. *Parking* = 0 is set as a reference.

[5] A duplex is an apartment with rooms on two adjoining floors connected by an inner staircase.

- *Garden* (Existence of garden), a dummy to control whether an apartment has a garden. *Garden* = 0 is set as a reference.
- *Basement* (Existence of basement), a dummy variable to control whether an apartment has a basement. *Basement* = 0 is set as a reference.
- *ExtraRoom* (Existence of extra room), a dummy variable to control whether an apartment has an extra room. *ExtraRoom* = 0 is set as a reference.
- *Terrace* (Existence of Terrace), a dummy variable to control whether an apartment has a terrace. *Terrace* = 0 is set as a reference.

Table III.5: Crossed table of Number of rooms and Number of bathrooms showing number of observations and correlation coefficient

	No. Bathrooms				
No. Rooms	0	1	2+	NA	Total
1	9,037	62,573	24	5,792	77,426
2	17,380	88,121	292	9,393	115,186
3	7,684	56,433	2,084	4,388	70,589
4	1,982	24,649	6,964	1,529	35,124
5	1,085	11,229	9,693	987	22,994
Total	37,168	243,005	19,057	22,089	321,319

	No.rooms	No.bathrooms
No.rooms	1	
No.bathrooms	0.3414	1

Correlation among some explanatory variables is another problem with hedonic regression because it can lead to inefficiency estimations. Apartments with more rooms should also have more bathrooms, and finding a one-room apartment with more than 2 bathrooms is almost impossible. Thus, number of rooms and number of bathrooms may be correlated, and *Stata* shows a correlation coefficient of 0.3408 (Table III.5). To avoid correlated regressors, a crossed variable of NbRoom and NbBath is created. The five dummies for number of rooms are crossed with the three dummies of number of bathrooms.

- *NbRoom/NbBath*, measured using 15 dummies: *P1S0* (1 room without bathroom), *P1S0, P1S1, P1S2, P2S0, P2S1, P2S2, P3S0, P3S1, P3S2, P4S0, P4S1, P4S2, P5S0, P5S1*, and *P5S2*. *P2S1* is removed as a reference.

Others variables that may be correlated are *Floor* and *Elevator* (Existence of elevator). The correlation coefficient for floor and existence of elevator is equal to 0.2477 (Table III.6). We consider that an elevator is necessary only for an apartment situated on the 4th floor and higher; in contrast, the existence of an elevator is not an interesting characteristic for a ground floor or a 1st floor apartment. *Floor/Elevator* is another crossed variable, only *Floor456* and *Floor7plus* dummies are crossed with three dummies for existence of elevator. The eight categories of *Floor/Elevator* are defined as follows.

- *F_Elev* (Floor/Elevator), measured by eight dummies: *L0* (ground floor), *L123*, *L456wElev* (4th, 5th, or 6th floor with an elevator), *L456w/oElev* (4th, 5th, or 6th floor without an elevator), *L456naElev* (4th, 5th, or 6th floor with unrecognized information on the existence of an elevator), *L7pluswElev*, *L7plusw/oElev*, and *L7plusnaElev*. *L123* is removed as a reference.

Table III.6: Crossed table of *Floor* and *Existence of Elevator* shows the number of observations and correlation coefficients

	Elevator		
Floor	N	O	Total
0	3,871	3,940	7,811
7	284	18,304	18,588
123	19,668	32,713	52,381
456	11,972	30,459	42,431
Total	35,795	85,416	121,211

	Floor	Elevator
Floor	1	
Elevator	0.2476	1

Semester is taken into account as a time control variable. This variable is included in the hedonic model to correct for trends in the market.

- *Semester*, measured using 20 dummies: respectively, *98S1* (1st semester of 1998), *98S2, 99S1, 99S2, 00S1, 00S2, 01S1, 01S2, 02S1, 02S2, 03S1, 03S2, 04S1, 04S2, 05S1, 05S2, 06S1, 06S2, 07S1,* and *07S2*. *00S1*, which is the most represented category, is removed as a reference.

In summary, for this first step of the study, price per square meter is regressed by these characteristic dummies. The estimated price obtained by this regression gives a value according to the utility of each characteristic compared with a reference characteristic. The reference apartment of our hedonic regression is defined as a standard 2-room apartment with 1 bathroom situated on the 1st, 2nd, or 3rd floor, built from 1948 to 1969 without parking, without garden, with basement, and without an extra room. This reference apartment is sold in the first semester of 2000. Details and frequency of regressors are presented in Table III.2.

4.2. Geostatistic methodology

We first define how to construct a semivariogram. Afterward, we explain our own method to examine a violation of the stationary assumption. Matheron (1963); Matheron (1965) defined geostatistics as spatial controls on variability. Observations taken in close proximity are more strongly correlated than observations separated by greater distance. Under *isotropic* and *spatial stationarity* assumptions, the degree of correlation can be defined as a function of distance.

Previous studies showed evidence of an anisotropic semivariogram (Besner (2002); Gillen, Thibodeau and Wachter (2001); Oden and Sokal (1986)). We decided not to examine the directional semivariogram because our data are divided into 36 rotating segments, and we do not have enough data to calculate a directional semivariogram. The directional semivariogram analysis may be intended for further research with additional data. Therefore, our objective is not to define an exact correlation structure, as an

isotropic correlation analysis is enough to show the violation of the stationary assumption.

The second assumption and the important one for our study is the stationary assumption. The second-order stationary assumption (or weakly stationary assumption) implies that the correlation function depends only on the distance between two locations. Let us define $s_i = (x_i, y_i)$ and $s_j = (x_j, y_j)$ to indicate geographic location of property i and j ((x_i, y_i) represents latitude and longitude coordinates) and $\varepsilon(S_i)$ and $\varepsilon(S_j)$ denote the hedonic residuals. The covariogram can be defined as a function of increment distance:

$$Cov\{\varepsilon(s_i), \varepsilon(s_j)\} = C(s_i - s_j) \equiv C(h) \qquad \text{Eq. III.2}$$

This covariogram can also be denoted as $C(h)$ with $h = s_i - s_j$ defining a distance between location s_i and s_j. A great-circle distance is calculated for our study. Other distance calculation methods exist, such as road distance, Euclidian distance, or exact time-based distance, but the great-circle distance is the most appropriate for our database. This empirical covariogram function tends to be high at small distances and falls with distance until a certain point, at which it becomes zero.

The semivariogram, denoted by $\gamma(h)$, is another useful function always applied to analyze spatial correlation structure. A semivariogram is defined as half the variance of the differences between all possible hedonic residuals separated by a constant distance h. The semivariogram function can be defined as follows (Eq. III.3):

$$\gamma(s_i - s_j) = \frac{1}{2} Var\{\varepsilon(s_i) - \varepsilon(s_j)\}$$
$$\gamma(h) = \frac{1}{2} E\{\varepsilon(s_i) - \varepsilon(s_j)\}^2 = C(0) - C(h) \qquad \text{Eq. III.3}$$

A semivariogram is an increasing function and becomes stable beyond a certain distance, $\gamma(h) \to C^* \; as \; h \to \infty$. The limited value of semivariogram C^* is called "*sill*" and the distance h_0, which makes $\gamma(h_0) \to C^*$, is called the "*range*" of the semivariogram. This "range" is the distance between two properties beyond which two

properties become uncorrelated. A semivariogram function is discontinuous near the origin, $\gamma(h) \to \theta_0 > 0 \; as \; h \to 0$, and the value of θ_0 is called a "*nugget*".

From our hedonic residuals, an empirical semivariogram is estimated using:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [\varepsilon(s_i + h) - \varepsilon(s_i)]^2 \qquad \text{Eq. III.4}$$

where $N(h)$ is the number of residuals locations separated by distance h. As the distance between two properties is a continuous value, calculating a semivariogram for all distances h is impossible because the empirical semivariogram is normally created for each distance interval. Our study estimates an empirical semivariogram for an interval of 50 meters, [0-50], [51-100], [101-150],...

The next step is to fit a functional form to the empirical semivariogram to estimate the range of the semivariogram necessary to compare the results of each segments and to examine a violation of the stationary assumption. Tu, Sun and Yu (2007) noted that identifying a true price spatial correlation structure is a challenge and determining whether the estimated correlation structure is the true spatial correlation is difficult. This paper estimates the spatial correlation using a selected semivariogram structure because our objective is not to identify a true spatial correlation structure but to show that the semivariogram analysis applied to housing data based on stationary hypothesis is inappropriate. Cressie (1991) proposed three theoretical semivariogram structures: a spherical function, an exponential function, and a Gaussian function. We chose only the spherical and the Gaussian semivariogram for our study. The spherical semivariogram is selected here because it provides a finite "range"; we use this estimated range to compare the semivariograms calculated from different segments. Gaussian semivariograms have no finite range but their function presents a very high correlation among observations separated by a very small distance. A comparison of the results estimated from these two models leads us to understanding the model that is a better to fit with our empirical semivariogram and shows a correlation gradient of Parisian property prices.

A spherical semivariogram is defined as follows:

$$\gamma(h;\theta) = \begin{cases} 0, & \|h\| = 0 \\ \theta_0 + \theta_1 \left[1{,}5 \left(\frac{\|h\|}{\theta_2} \right) - 0{,}5 \left(\frac{\|h\|}{\theta_2} \right)^3 \right], & 0 < \|h\| \leq \theta_2 \\ \theta_0 + \theta_1, & \|h\| \geq \theta_2 \end{cases} \quad \text{Eq. III.5}$$

A Gaussian semivariogram is defined as follows:

$$\gamma(h;\theta) = \begin{cases} 0, & \|h\| = 0 \\ \theta_0 + \theta_1 \left(1 - exp - \left(\frac{\|h\|}{\theta_2} \right)^2 \right), & \|h\| > 0 \end{cases} \quad \text{Eq. III.6}$$

The *nugget* is θ_0, the *sill* is $\theta_0 + \theta_1$, and the *range* is θ_2.

The *Gstat* package in *R* is used to fit the theoretical semivariogram to the empirical one using a relatively simple method, the weighted least squares. We need to initialize the value of a nugget and a sill. The nugget is initialized to the average of the first three estimated semivariograms (from three intervals [0–50], [51–100], and [101–150]). The sill is initialized to the average of the variogram obtained from a distance greater than 2 kilometers and the initialization of the range is 500 meters. The maximum distance used to calculate the semivariogram between two residuals is set to 3 kilometers.

4.3. Stationary assumption analysis

Two kinds of stationary assumptions are examined in this paper: the time stationary assumption and the spatial stationarity assumption. For the first step, the time stationary assumption is examined by comparing semivariograms obtained from data during different periods. A fitted semivariogram shape and an estimated "range" among our 10 years of data should be similar under the time stationary assumption. Moreover, these results obtained from whole data (called a 10-year semivariogram) should also be similar to the results obtained from annual data (called a 1-year semivariogram).

The second step concerns the main point of our study, and we examine the violation of the spatial stationarity assumption by comparing the semivariogram obtained from different data segments. Under the spatial stationarity assumption, the

semivariogram should be constant across space. In the particular case for which the local experimental semivariogram varies with spatial location, a stationary variogram model may be inappropriate (Atkinson and Lloyd (2007)). As previously noted, the data are divided into 36 rotating segments around the *Arc de Triomphe* and the *Place d'Italie*. For the data located within each segment, a local spatial stationarity assumption is assumed to calculate an experimental semivariogram but a global stationarity assumption is not imposed. The shape of the experimental semivariogram obtained from each of the 36 segments and the estimated range among these 36 segments are compared. Our testing hypothesis is: under a spatial stationarity assumption, the empirical semivariograms obtained from different segments should be similar. Moreover, the estimated range of the semivariograms obtained from each segment (called a local semivariogram) should have the same magnitude and should be similar to the result obtained from whole data analysis (called a global semivariogram). If the semivariogram varies by segment, then the geostatistic approach applied to all data under a global spatial stationarity assumption is not appropriate.

5. Estimated ranges and stationary analysis

Our results are described in three parts according to our three step methods, hedonic regression results, semivariogram obtained from the data of different period and semivariogram obtained for the different data segment.

5.1. Hedonic regression

Our regression analysis includes 35 dummy variables for the physical characteristics and 19 dummy variables for semester. This time variable helps capture the value of external factors that can indirectly affect property transaction prices other than their own characteristics, i.e., global market trends, evolution of interest rates, or evolution of exchange rates.

Table III.7: OLS estimation results (without spatial characteristics)

Variable	Estimated coefficient	Std. Dev.	t-statistics	p-value
Period==1850_1913	-12.38	7.71	(-1.61)	0.108
Period==1914_1947	-41.47***	8.77	(-4.73)	0.000
Period==1970_1980	-223.41***	9.57	(-23.34)	0.000
Period==1981_2000	359.08***	12.47	(28.80)	0.000
Period==Af2001	911.32***	23.28	(39.15)	0.000
Period==Bf1850	837.60***	12.66	(66.17)	0.000
Period==NA	132.57***	9.89	(13.41)	0.000
Parking==1	159.63***	7.98	(20.02)	0.000
ExtraRoom==1	876.11***	13.25	(66.11)	0.000
AprtTyp==AprtDu	548.37***	14.86	(36.89)	0.000
L_Elev==L0	-297.96***	9.02	(-33.04)	0.000
L_Elev==L456naElev	105.43***	6.02	(17.52)	0.000
L_Elev==L456w/oElev	-154.02***	12.86	(-11.98)	0.000
L_Elev==L456wElev	259.16***	8.39	(30.89)	0.000
L_Elev==L7plusnaElev	132.65***	16.45	(8.07)	0.000
L_Elev==L7plusw/oE~v	-113.02	78.98	(-1.43)	0.152
L_Elev==L7pluswElev	-28.73**	10.87	(-2.64)	0.008
NbRoom_NbBath==p1s0	-368.02***	14.29	(-25.75)	0.000
NbRoom_NbBath==p1s1	69.36***	6.89	(10.07)	0.000
NbRoom_NbBath==p1s2	-116.37	259.43	(-0.45)	0.654
NbRoom_NbBath==p2s0	-478.07***	10.80	(-44.27)	0.000
NbRoom_NbBath==p2s2	717.30***	74.69	(9.60)	0.000
NbRoom_NbBath==p3s0	-272.87***	15.31	(-17.82)	0.000
NbRoom_NbBath==p3s1	117.71***	6.91	(17.04)	0.000
NbRoom_NbBath==p3s2	1147.88***	28.33	(40.52)	0.000
NbRoom_NbBath==p4s0	-18.50	28.97	(-0.64)	0.523
NbRoom_NbBath==p4s1	286.50***	9.32	(30.73)	0.000
NbRoom_NbBath==p4s2	615.09***	16.25	(37.86)	0.000
NbRoom_NbBath==p5s0	59.07	38.99	(1.52)	0.130
NbRoom_NbBath==p5s1	568.93***	13.34	(42.65)	0.000
NbRoom_NbBath==p5s2	854.43***	14.26	(59.93)	0.000
Terrace==1	739.47***	14.66	(50.45)	0.000
Garden==1	435.30***	28.44	(15.30)	0.000
Basement==1	12.92*	5.79	(2.23)	0.026
Semester==00s2	153.23***	14.25	(10.75)	0.000
Semester==01s1	245.60***	14.32	(17.15)	0.000
Semester==01s2	358.53***	14.35	(24.98)	0.000

Variable	Estimated coefficient	Std. Dev.	t-statistics	p-value
Semester==02s1	452.89***	14.36	(31.53)	0.000
Semester==02s2	645.84***	14.45	(44.71)	0.000
Semester==03s1	864.50***	14.97	(57.74)	0.000
Semester==03s2	1121.89***	14.69	(76.37)	0.000
Semester==04s1	1374.62***	14.42	(95.35)	0.000
Semester==04s2	1666.86***	14.27	(116.79)	0.000
Semester==05s1	2031.65***	14.35	(141.58)	0.000
Semester==05s2	2438.76***	14.45	(168.69)	0.000
Semester==06s1	2660.15***	14.57	(182.60)	0.000
Semester==06s2	2922.72***	14.73	(198.41)	0.000
Semester==07s1	3153.54***	14.70	(214.46)	0.000
Semester==07s2	3460.80***	14.78	(234.19)	0.000
Semester==98s1	-459.83***	14.94	(-30.77)	0.000
Semester==98s2	-395.42***	14.38	(-27.51)	0.000
Semester==99s1	-311.31***	14.02	(-22.21)	0.000
Semester==99s2	-146.18***	13.79	(-10.60)	0.000
_cons	2634.30***	13.46	(195.73)	0.000
r2	0.539			
N	298891			

Dependent variable: price per square meter. The symbols *, **, and *** denote significance at the 5%, 1%, and 0.01% levels, respectively. The reference category is a standard 2-room apartment with 1 bathroom situated on the 1st, 2nd, or 3rd floor, built from 1948 to 1969, without parking, without a garden, with a basement, and without an extra room; this reference apartment is sold in the first semester of 2000.

Table III.7 shows our hedonic regression results of price per square meter with apartment characteristics. Our $R^2 = 53.88\%$; which is quite small compared with other previous studies that always show R^2 greater than 80%. We take into account only physical and time characteristics. Spatial characteristics, which normally have significant explanatory power, are not included in this hedonic regression. Another reason for the low R^2 is that our dependent variable is price per square meter. In some previous papers, price is considered a dependent variable and surface is taken into account as a regressor. Because surface is a principal explanatory variable of transaction price, the regression results evidently provide a higher value of R^2. If we replace our dependent variable with price and surface is considered as an explanatory variable, we find $R^2 = 82.61\%$ and $R^2 = 79.59\%$ for a regression with and without spatial characteristics, respectively.

The results from the hedonic regression are consistent with our expectations. Almost all regression coefficients are significant at the 5% level, except for some estimators for the construction period 1850–1913, for apartments situated on the 7th floor and higher and without an elevator (*7plusw/oElev*), and for three crossed variable number of rooms/number of bathrooms (*P1S2*, *P4S0* and *P5S0*). This regression coefficient view in Table III.7 represents the estimated value of each characteristic compared with the reference characteristic. We recall that the reference apartment of our hedonic regression is defined as a standard 2-room apartment with 1 bathroom situated on the 1st, 2nd, or 3rd floor, built during 1948 to 1969 without parking, without a garden, with a basement, and without an extra room, and this reference apartment is sold in the first semester of 2000.

One of our principal characteristics is the number of rooms and the number of bathrooms. The coefficients for this crossed variable *NbRoom/NbBath* are all significant, except for the three coefficients for *P1S2*, *P4S0*, and *P5S0*. These non-significant estimators are explained by a few numbers of observations in these three categories. Seldom does a one-room apartment have 2 bathrooms or does a four-room apartment have no bathroom. Our regression result shows that the price of a one-room apartment with 1 bathroom (*P1S1*), called a *studio*, is €69 per square meter higher than the price of a 2-room apartment with 1 bathroom (*P2S1*) which is our reference category. This result confirms what we observe in the Paris property market. A studio is the most wanted apartment for students and young workers, and rent per month per square meter for a studio is always higher than other apartment types. Another interesting result is that the existence of a bathroom is an important characteristic because the price of an apartment with 1, 2, or 3 rooms but without a bathroom is significantly lower than the same type of apartment with 1 or more bathrooms.

The crossed variable *Floor*/*Elevator* gives us other interesting results and confirms what we observe in the Paris real estate market. Ground floor apartments cost €297 per square meter less than a 1st, 2nd, or 3rd floor apartment. The price is higher when we move to a 4th, 5th, or 6th floor apartment only if the building has an elevator. For buildings with no elevator, moving from the 1st, 2nd, or 3rd floor to a higher floor reduces significantly price per square meter. We find a negative coefficient for the 7th floor apartment (*F7pluswElev* and *7plusw/oElev*). Given that bourgeois buildings in Paris

are Haussmannian style, this building normally has seven floors. On the top floor are maids' bedrooms (*chambre de bonne*), small rooms with a small bathroom or sometimes without a bathroom, and only accessible from a separate staircase behind building. Thus, moving to the 7th floor reduces the price per square meter.

Compared with the reference apartment constructed during 1948 to 1969, an apartment constructed after 1980 and the most ancient apartment constructed before 1850 has a higher price. The estimator of the period 1850–1913 is not significant, indicating that no significant price difference exists between an apartment constructed during this period and a reference period.

The existence of extra room, parking, a terrace, a garden, or a basement increases price per square meter. In particular, the existence of a terrace or garden increase price per square meter by approximately €739 and €435, respectively, and a duplex apartment is more expensive than a standard one.

Table III.8: Descriptive Statistics of Residuals

Variable	Obs	Mean	Std. Dev.	Min	Max
Residu_1	298,891	13.4	1,269.1	-5,767.1	16,615.4
residu2007	27,834	29.0	1,648.7	-5,991.6	13,232.4
residu2006	28,502	25.9	1,500.5	-5,337.2	13,767.4
residu2005	30,340	24.1	1,439.5	-5,720.3	12,750.4
residu2004	30,783	16.0	1,351.7	-4,787.3	15,194.4
residu2003	27,106	11.8	1,265.2	-4,549.7	14,162.2
residu2002	29,869	8.9	1,184.9	-4,120.9	16,559.9
residu2001	30,401	6.9	1,168.1	-4,153.0	15,421.5
residu2000	31,780	6.6	1,075.7	-3,609.1	14,061.1
residu1999	34,298	4.2	944.4	-3,271.3	16,774.4
residu1998	27,978	2.3	850.7	-2,745.6	9,503.6

Table III.8 shows the descriptive statistics of the residuals obtained from an annual regression. *Residu_1* represents the residual obtained from a 10-year data regression. Considering these residual results, mean values are near zero with standard deviations from €850 to €1,648 per square meter. High standard deviation values can confirm that transaction prices cannot be explained only by physical characteristics and that spatial characteristics are important explanatory variables.

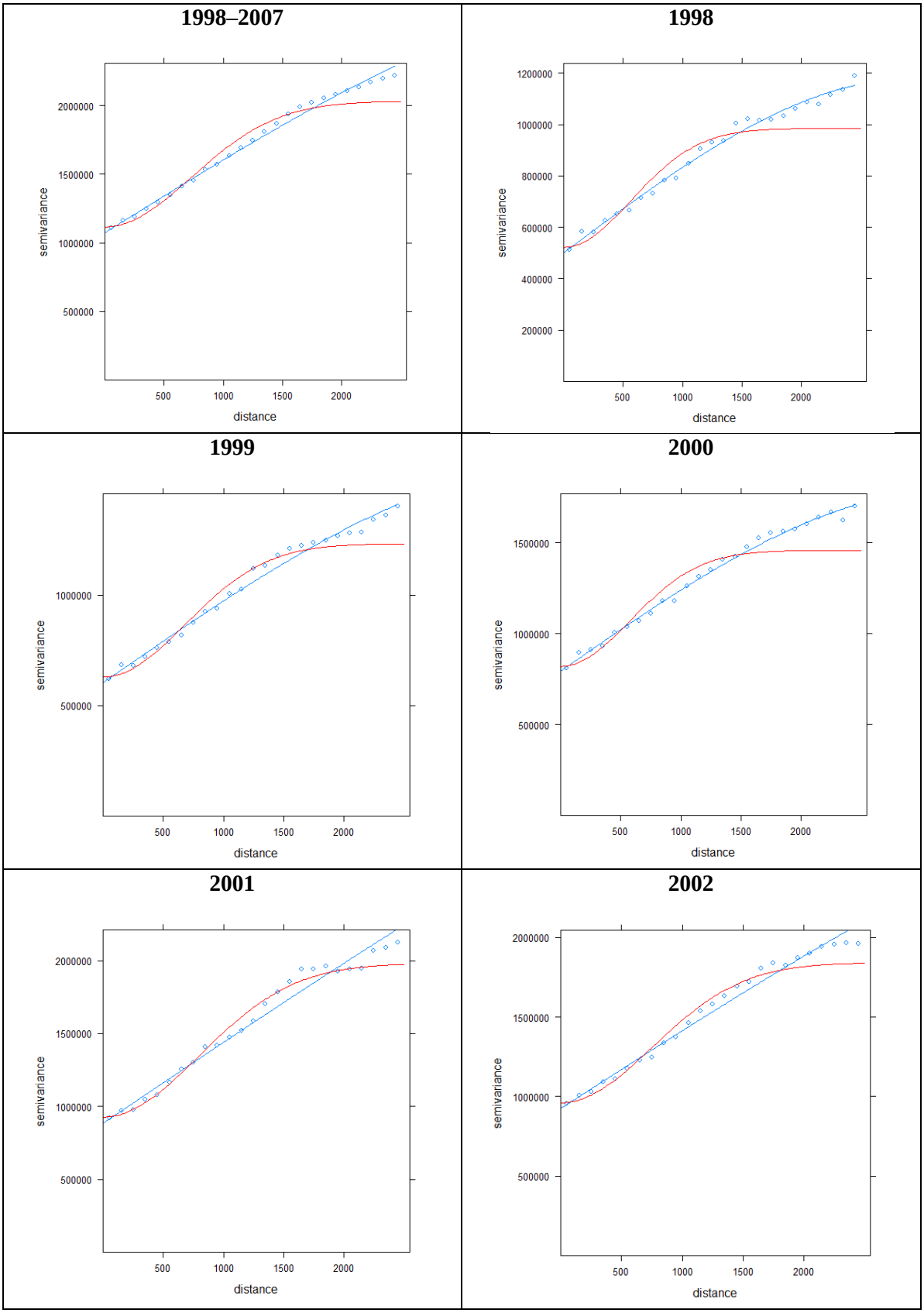
5.2. Stationary semivariogram analysis

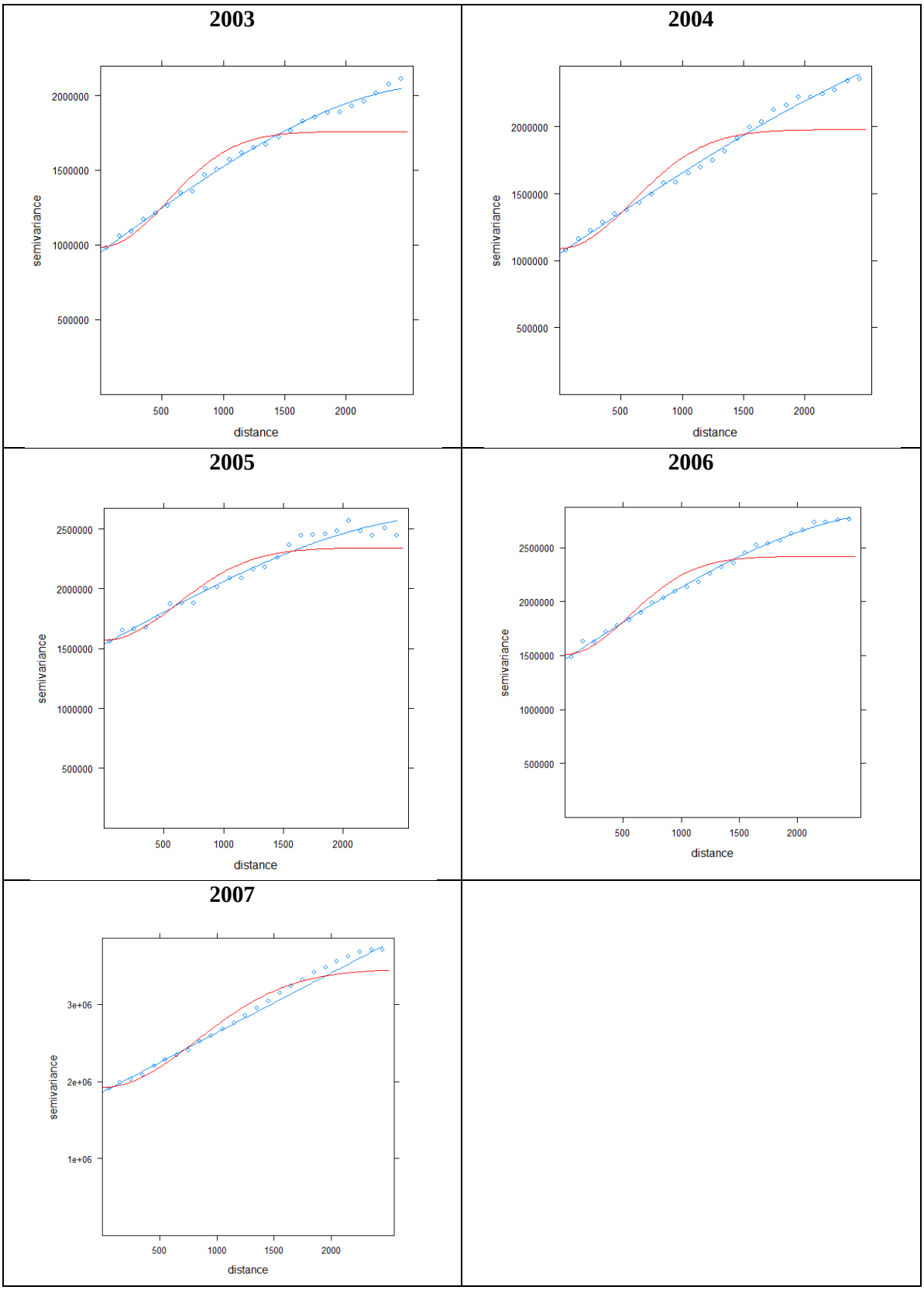
We first estimate a 10-year semivariogram to obtain a global view of property price correlation for 10 years of data for all of the data segments around the *Arc de Triomphe*. The estimated range of this 10-year semivariogram is 1,032 meters, which means that, on average, apartment prices around the *Arc de Triomphe* are uncorrelated if the distance separating them is greater than 1,032 meters. We are then interested in the variation of an experimental semivariogram over a study period and the variation of an experimental semivariogram over a study area.

5.2.1. Time stationary analysis

If each 1-year semivariogram has a similar shape and a similar estimated range, we assume that the spatial process is stationary in time. Figure III.3 shows 10 semivariograms obtained from the one- year data.

Figure III.3 : 10-year semivariogram and 1-year semivariogram



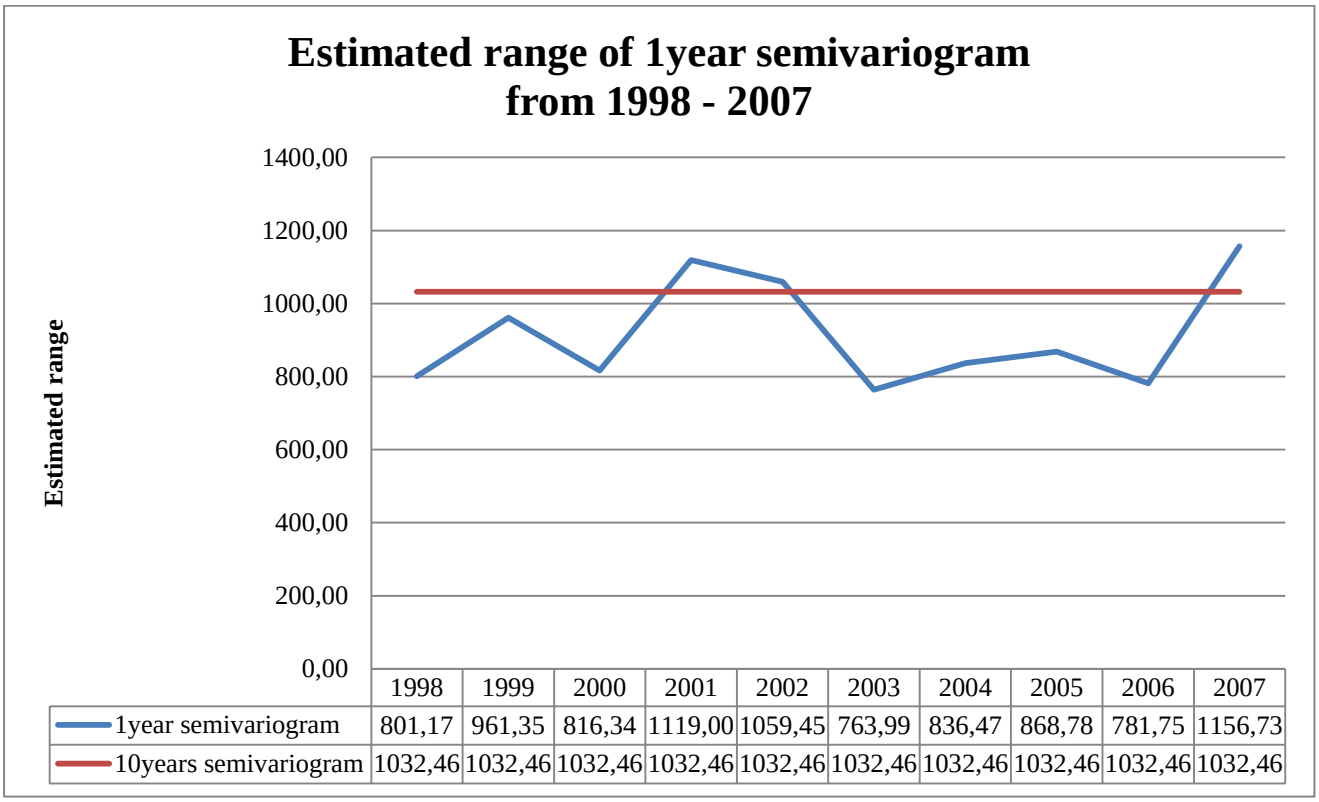


Scatter plot: Empirical semivariograms; blue line: Spherical fitted semivariogram and red line: Gaussian fitted semivariogram.

We observe that these 10 semivariograms have the same linear shape; therefore, estimating range parameters is very difficult. However, a spherical model (blue line) and a Gaussian model (red line) are chosen to fit these experimental semivariograms, but a spherical model cannot provide a valid estimated range. From the Gaussian model, we obtain a variability of estimated ranges over the study period and these 1-year semivariogram estimated ranges are also different from the 10-year range (Figure III.4). The estimated range can describe how the data are spatially dependent. A long range indicates high correlation and a short range indicates low correlation. Parisian residential prices are highly correlated in 2001, 2002, and 2007 and have low correlation in 2003 and 2006. During 10 years, semivariogram autocorrelation ranges vary from 781 meters to 1,156 kilometers. In 1998, properties around the *Arc de Triomphe* are correlated with respect to price if the distance between them is less than 801 meters; however, in 2007, property prices become uncorrelated only if the distance between them is more than 1,156 kilometers. Nevertheless, our results cannot really confirm whether the temporal structure of our spatial data is stationary. We observe the same shape for the semivariograms but the estimated ranges are different.

We attempt to explain these different ranges over time for two reasons. First, real estate is an illiquid asset and is not traded every year. Therefore, spatial distribution is different from one year to another and this may cause variability in the spatial structure over time. The second reason is that certain economic factors may have an effect on property prices, such as interest rates, tax rates, or capital growth; however, these factors are not included in our hedonic regression. A change in the year-over-year factors may cause variability in the spatial structure.

Figure III.4: Estimated range for 1-year and 10-year semivariograms



	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
1year semivariogram	801,17	961,35	816,34	1119,00	1059,45	763,99	836,47	868,78	781,75	1156,73
10years semivariogram	1032,46	1032,46	1032,46	1032,46	1032,46	1032,46	1032,46	1032,46	1032,46	1032,46

5.2.2. Spatial stationarity analysis

To examine a violation of spatial stationarity assumption, we split our data into 90° rotating windows around the *Arc de Triomphe* and the *Place d'Italie*, and we examine a semivariogram obtained from these different segments. If the semivariogram varies locally among our 36 regions of interest, the spatial process cannot be assumed as stationary. We find that our semivariograms vary dramatically across the region of interest. Only the semivariogram of six segments around the *Arc de Triomphe* is presented here: [350° – 80°], [20° – 110°], [80° – 170°], [150° – 240°], [180° – 270°], and [240° – 330°]. Other results are showed in the Polar chart (Figure III.11).

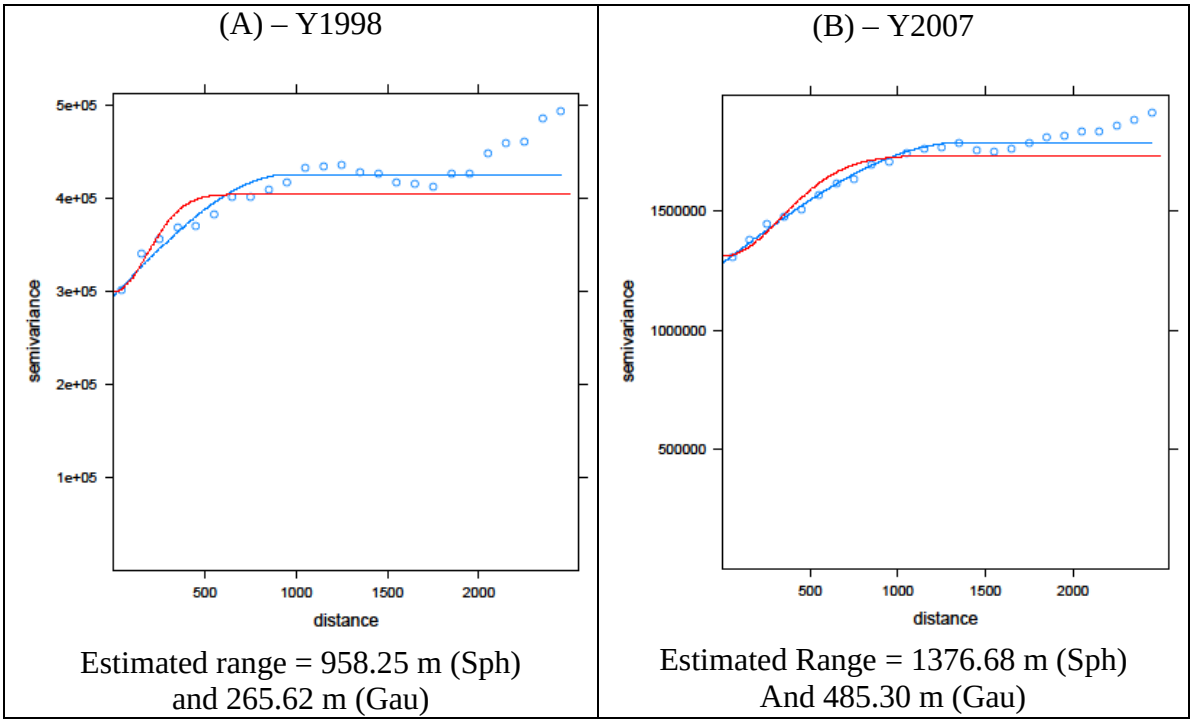
[350° – 80°]: 17th Arrondissement

Figure III.5 clearly shows that the spherical model fits our empirical semivariogram better than the Gaussian model. We keep here only an estimated range obtained from a spherical semivariogram. For 2007, the estimated range is 1.38 kilometers, which is approximately 400 meters longer than the 958 meters estimated range for 1998. Segment [350° – 80°] takes into account properties situated in the 17th arrondissement. This estimated range shows that prices of residential properties in the 17th arrondissement are correlated if the distance between them is less than 1.38 kilometers and this correlation declines if the distance between properties is greater than 1.38 kilometers. We also observed that correlation increases again if the distance is greater than 1.8 kilometers. This increasing variance may be the result of less data because only Parisian residential transaction prices are included in our database.

We attempt to explain this estimated range using the specific characteristics of the 17th arrondissement residential market. This arrondissement presents a combined characteristic of the 16th arrondissement, which is a bourgeois residential area, and the 18th arrondissement, which is a middle-class residential area. The rail network from *Gare Saint Lazare* to the eastern Parisian suburbs is located approximately 2.25 kilometers from the *Arc de Triomphe* in the direction of 35 degrees, and is the border for this north–south market segmentation. Residential property prices in the 17th arrondissement are not uniform; great price variation exists between the north area and the south area. Property prices are higher in the south area near the *Arc de Triomphe*, called *Terne*, and the *Plaine-de-Monceau* district with its high-end residential and commercial area. With traditional architecture Haussmannian buildings and large boulevards, the resident market in the south area of the 17th arrondissement provides an elegant feeling. Moreover, foreign investors and French bourgeois clients are interested in this area. The north area (*Epinnette* and *Batignolles* districts) is livelier and has a local community, and is sometimes considered a working-class district. The north area of the 17th arrondissement is near the 18th arrondissement, known for its bustling immigrant neighborhoods with high criminal activity. This price variation between the north area and the south area of the 17th arrondissement may explain the estimated range of 1.38 kilometers. Property prices in the south area near the *Arc de Triomphe* are not correlated with property prices

in the *Epinnette* district because the distance between these two areas is greater than 1.38 kilometers.

Figure III.5: Semivariogram of window [350° – 80°]: 17th arrondissement



Scatter plot: Empirical semivariogram; blue line: Spherical fitted semivariogram (Sph) and red line: Gaussian fitted semivariogram (Gau).

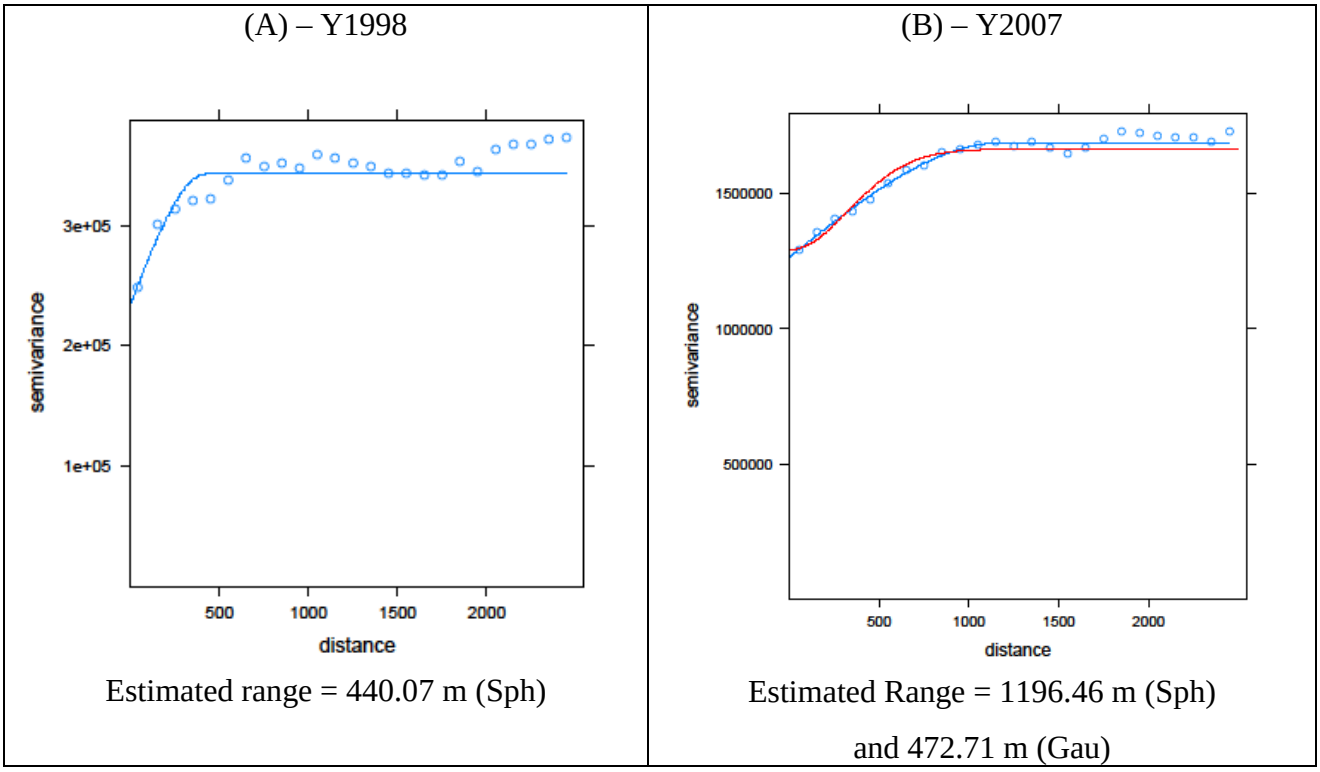
[20° – 110°]: 8th Arrondissement

Again, for the segment [20° – 110°], the spherical model fit the experimental semivariogram better than the Gaussian model. An estimated range is 440 meters in 1998 and 1.20 kilometers in 2007. A comparison of the estimated range obtained from this segment and the results of the first segment showed values that are very different for 1998 but that are less different in 2007. Between 1998 and 2007, an increasing range shows the development of the real estate market in this area, and the correlation among property prices increased.

This segment that takes into account properties located in the 8th arrondissement and the east area of the 17th arrondissement. The east area of the 17th arrondissement is the area around *Parc de Monceau*, an aristocratic park surrounded by many wealthy

hotels and bourgeois buildings. Moreover, around *Parc de Monceau* is one of the most expensive resident areas of the 17th arrondissement. The 8th arrondissement, a part of the "Golden Triangle" (*Triangle d'Or*), is both a central business district and a resident area for the high bourgeoisie class. These two expensive areas of Paris evidently have property prices that are correlated with each other. Our results confirm that, in 2007, prices of residences located to the east of the 17th arrondissement are correlated with prices of residences located in the 8th arrondissement if the distance between them is less than 1.20 kilometers. These results also show that property prices for different administrative boundaries may be correlated.

Figure III.6: Semivariogram of window [20° – 110°]: Parc de Monceau



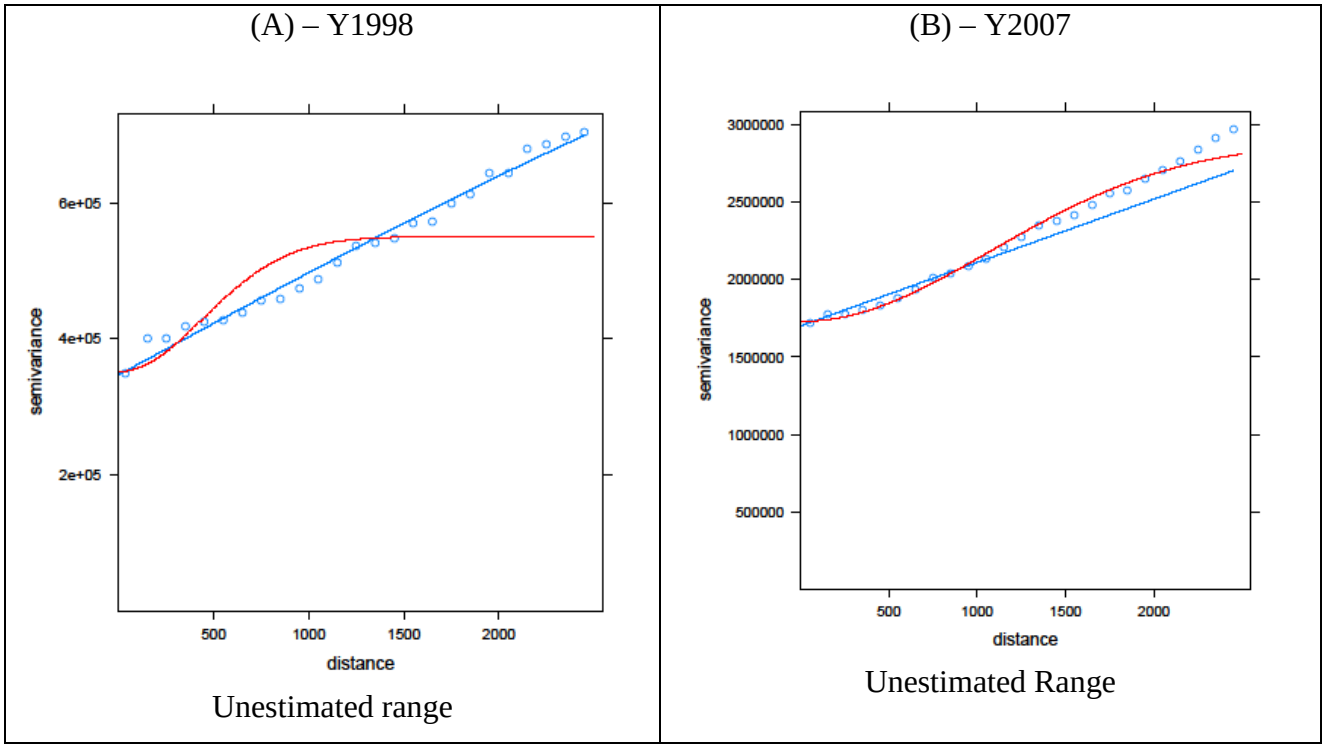
Scatter plot: Empirical semivariogram; blue line: Spherical fitted semivariogram (Sph) and red line: Gaussian fitted semivariogram (Gau).

[80° – 170°]: Avenue des Champs-Elysées

The window [80° – 170°] takes into account the properties located throughout and around the famous *Avenue des Champs-Elysées*. *Avenue des Champs-Elysées* is a luxury avenue starting from the *Arc de Triomphe* and going to the *Place de la Concorde*, for a

total distance of 2.2 kilometers. Ours results show that residential property prices along the *Avenue des Champs-Elysées* are correlated. For 1998 and 2007, semivariograms are increasing functions and neither of the two theoretical semivariograms provides a valid estimated range.

Figure III.7: Semivariogram of window [80° – 170°]: Avenue des Champs-Elysées



Scatter plot: Empirical semivariogram; blue line: Spherical fitted semivariogram (Sph) and red line: Gaussian fitted semivariogram (Gau)

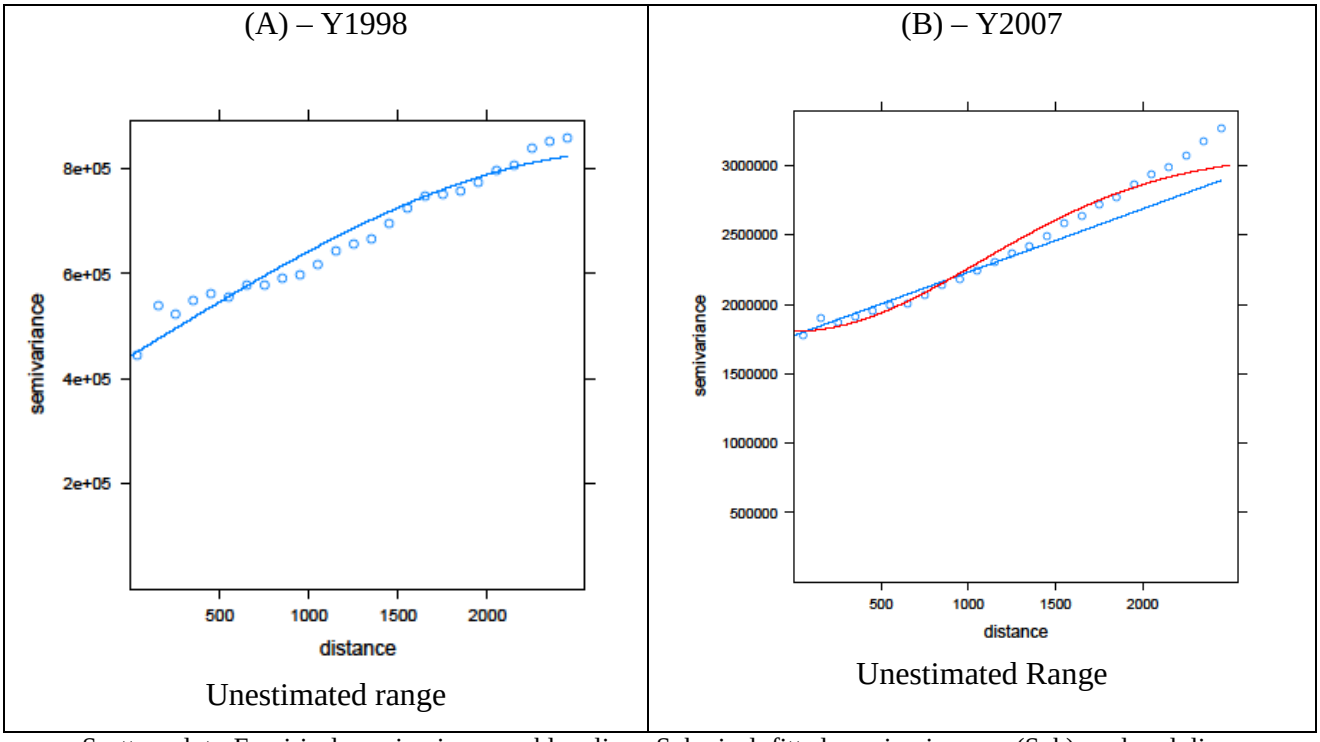
[150° – 240°]: Eiffel Tower

Similar to the prior results for *Avenue des Champs-Elysées*, the semivariogram estimated from segment [150° – 240°] is an increasing semivariogram without any valid estimated range. The correlation among property prices declines when the distance separating two properties increases but never becomes zero. We cannot estimate the distance beyond which the semivariogram is stable because it is a non-transitive[6] linear semivariogram as a result of the existence of many small markets and the correlation for

[6] The semivariogram is transitive if it reaches a finite sill at a finite range (R. Olea, *"Stochastic Modeling and Geostatistics: Principles, Methods and Cases Studies*," Ed. J. Yarus and R. Chambers, 1994).

each small market, which are linked together into one large resident market around the *Arc de Triomphe* and the Eiffel Tower.

Figure III.8: Semivariogram of window [150° – 240°]: Eiffel Tower



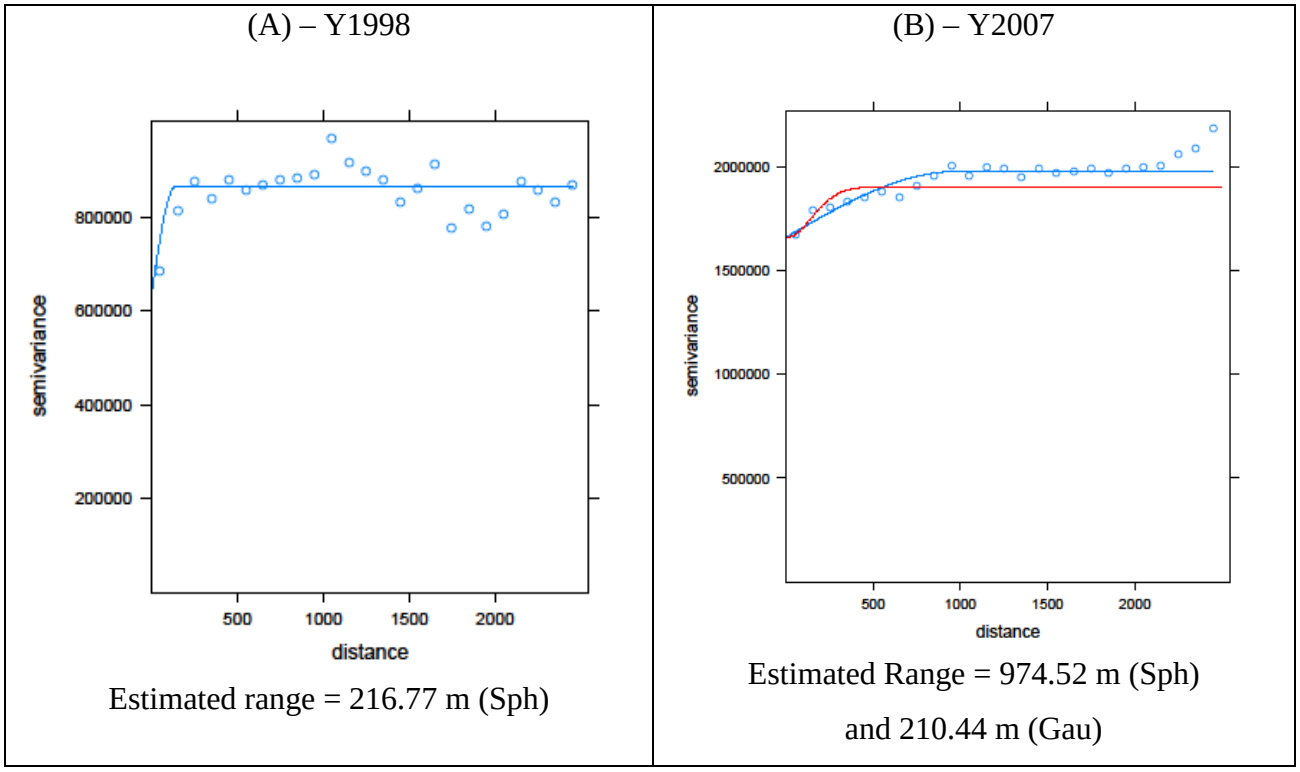
Scatter plot: Empirical semivariogram; blue line: Spherical fitted semivariogram (Sph) and red line: Gaussian fitted semivariogram (Gau)

[180° – 270 °]: 16th arrondissement (Avenue Victor Hugo)

For the segment [180° – 270°], the estimated range is shorter than others previous segments, at only 216 meters in 1998 and 974 meters in 2007. These results show that, in 1998, only prices for property located close to each other are correlated; however, in 2007, property prices were correlated if the distance between them was less than 974 meters. This data segment includes all properties in the 16th arrondissement. The residential market of the 16th arrondissement shows the same characteristics as the residential market of the 17th arrondissement. The district can be divided in two parts: the north part (*Torcadero*, *Passy*, and *la Muette* districts) and the south part (*Auteuil* district). These two parts are bounded by *rue d'Auteuil*. The north area is popular as a bourgeois residential area bordered by prestigious avenues. The area has many attractive

advantages: building quality, green environment with the presence of *Ranelagh* Gardens, shopping street, and reputed schools. This area attracts investors, foreign clients, and French bourgeois families. The south area is a working class residential area with noise pollution from traffic and buildings with a heterogeneous quality, which is why the residential market south of the 16th arrondissement is not correlated with the north area. Our results confirm that property prices in the north area near the *Arc de Triomphe* are not correlated with property prices in the south area, the *Auteuil* district, because the distance between them is certainly greater than 974 meters.

Figure III.9: Semivariogram of window [180° – 270 °]: 16th arrondissement (Avenue Victor Hugo)



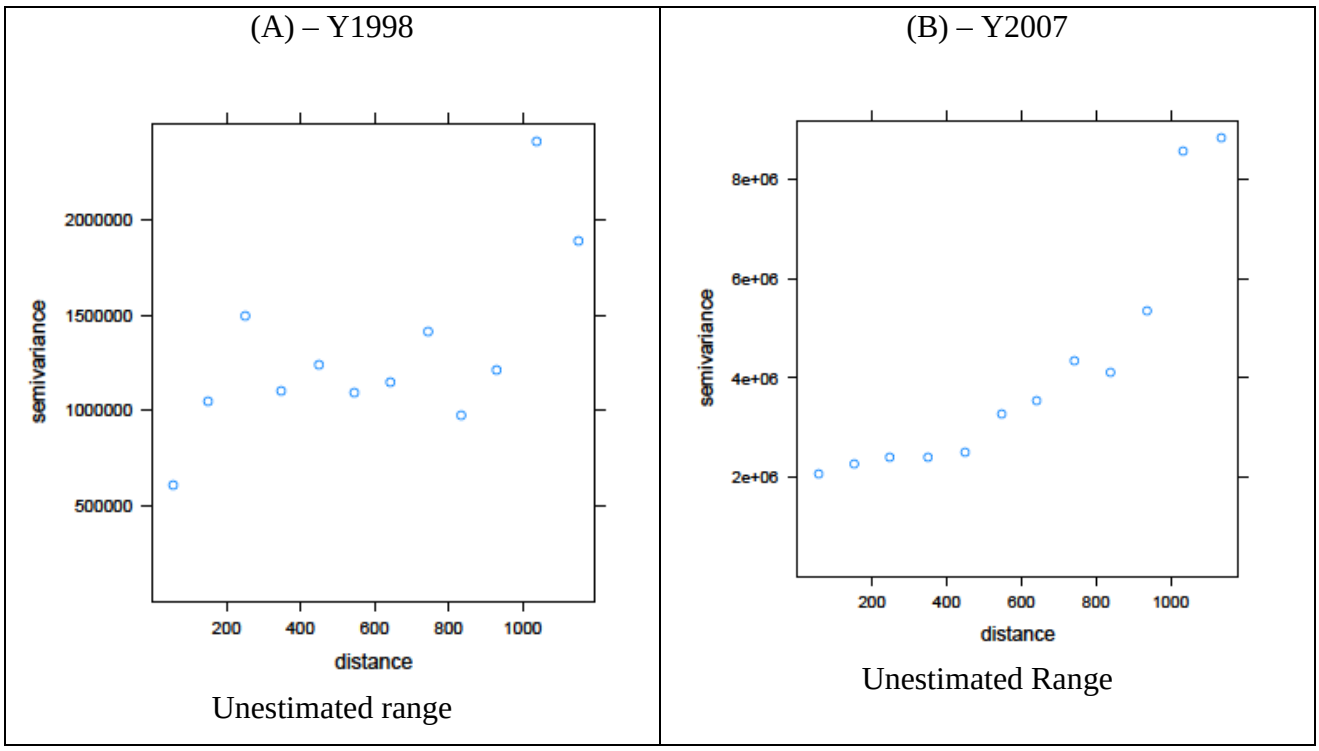
Scatter plot: Empirical semivariogram; blue line: Spherical fitted semivariogram (Sph) and red line: Gaussian fitted semivariogram (Gau)

[240°– 330°]: Palais de Congrès

The results of the [240°– 330°] window present a particular graphic showing that the theoretical semivariogram cannot be estimated. These results are different from what was obtained from the [80° – 170°]. For this segment, the unestimated range is caused by

an insufficient sample in this segment, which takes into account the data from the northeast direction from the *Arc de Triomphe* to the edge of Paris. This segment shows that, in the case of a few observations, estimated properties price correlations using semivariogram analysis is not the most appropriated one.

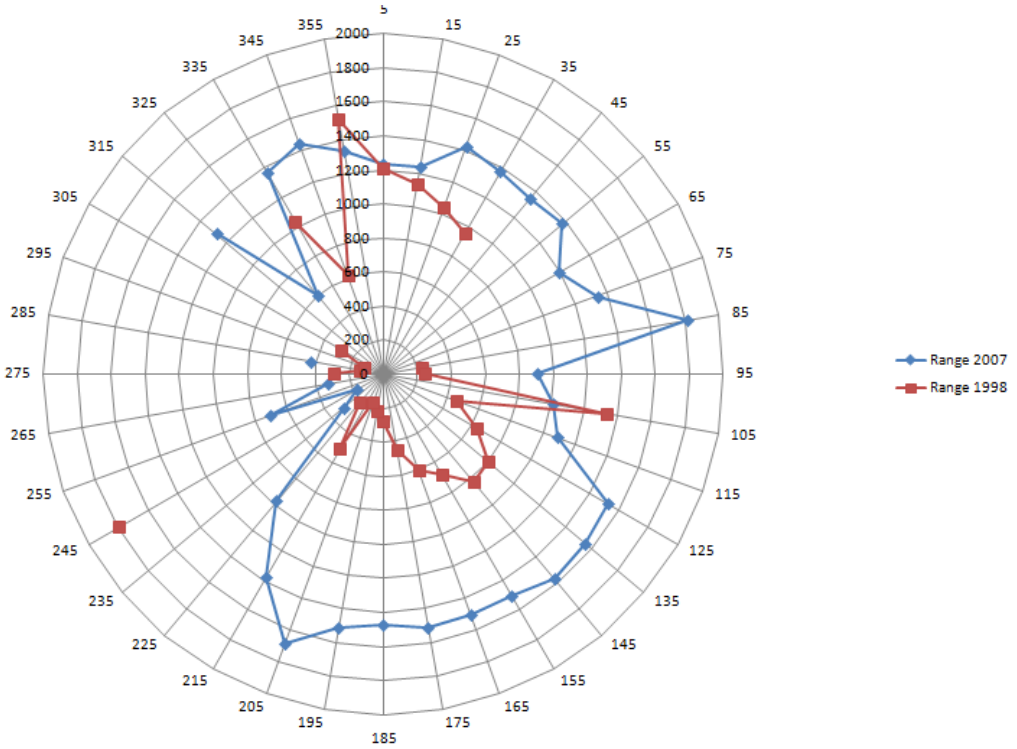
Figure III.10: Semivariogram of window [240°– 330°]: Palais de Congrès



Scatter plot: Empirical semivariogram; blue line: Spherical fitted semivariogram (Sph) and red line: Gaussian fitted semivariogram (Gau).

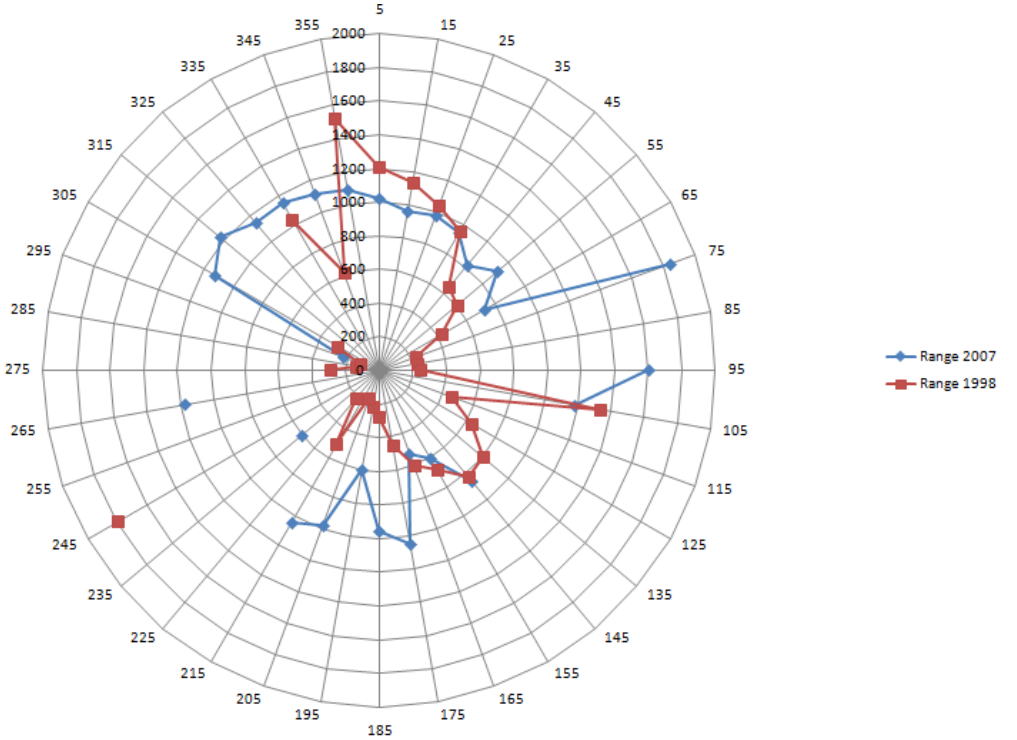
The results obtained from 36 data segments around the *Arc de Triomphe* show that the estimated semivariograms vary dramatically across the region of interest. When we look in detail at each segment (six segments are chosen here), the results obtained from the semivariogram analysis depend deeply on the structure of the residential market in each segment. Figure III.11 and Figure III.12 show the polar chart for the estimated ranges obtained from 36 segments around the *Arc de Triomphe* and the *Place d'Italie*. The variation in estimated ranges around the *Place d'Italie* also confirms our idea of non-spatial stationarity data. The natural park, rail network, prestigious boulevard are considered as market segmentation borders and affect the variations in the semivariogram results (form and estimated range). The global stationary assumption cannot be automatically assumed to apply to the geostatistic model.

Figure III.11: Polar chart of 36 estimated ranges around the *Arc de Triomphe*, 1998 and 2007



Blue line: Estimated ranges of semivariograms from 36 data segments around the *Arc de Triomphe* in 2007
Red line: Estimated ranges of semivariograms from 36 data segments around the *Arc de Triomphe* in 1998

Figure III.12: Polar chart of 36 estimated ranges around the *Place d'Italie*, 1998 and 2007



Blue line: Estimated ranges of semivariogram from 36 data segments around the *Place d'Italie* in 2007
Red line: Estimated ranges of semivariogram from 36 data segments around the *Place d'Italie* in 1998

The logarithmic and square root transformations are applied to the residuals, and this data deformation can reduce the heterogeneity of the semivariogram shapes and estimated ranges. However, we cannot obtain a homogenous estimated range.

6. Semivariogram range sensitivity analysis

As the empirical results show, the semivariogram range varies among the data segments. Our next question is, "what should be a cause of the variation in the range?" This section aims to analyze the factors causing variations in the range and to study the sensibility of the variogram range.

Our first attempt is to apply the logarithmic transformation proposed by Kerry and Oliver (2007a) that allows reduced heterogeneity of the semivariogram shape. Despite several attempts at data transformation (square, square root, logarithmic), we cannot obtain total homogenous semivariograms.

Our second attempt is to find property characteristics that may cause this variability in the semivariograms. By including certain explanatory variables in the hedonic regression, the variogram range for each segment is then re-estimated. If the new variogram range becomes less heterogeneous after including an explanatory variable in the hedonic model, we can conclude that such explanatory variables may cause the non-stationary problem. Identifying the explanatory variable that causes variability in the estimated range is a necessary step in understanding spatial correlation patterns and obtaining stationary estimated ranges, and is an important step in geostatistic analysis. This estimated range is used to determine the distance until which the correlation among properties prices disappears to define the real estate submarket (Tu, Sun and Yu (2007)), or is used in property price prediction (Basu and Thibodeau (1998)).

Four variables are chosen for inclusion in this study of sensitivity: two socio-demographic variables and two submarket variables. The two socio-demographic variables chosen are *age of buyer* and *age of seller*. We observe that same-age people tend to live in the same area. Older people tend to leave the town center and move outside

of the town; inversely, young people seeking to reduce their transportation costs and time tend to move into the center of towns. Consequently, socio-demographics may be one omitted variable that causes variability in the variogram range. Arrondissement, an indicator of administrative boundary, is a submarket explanatory variable chosen. As showed in Bourassa, Cantoni and Hoesli (2007), submarket variables added to the hedonic model can improve the accuracy of housing price predictions by the geostatistic model. Another submarket variable is the quarter index. Paris is dividend in 20 arrondissements and each arrondissement is dividend in four quarters; therefore, 80 indices of the quarters are included in this study. The result obtained for the hedonic model including number of quarters is compared with the result obtained with number of arrondissement, which enables the conclusion that including such an accurate spatial variable may better capture spatial variability and lead to a stable estimated variogram range.

Because we noted that neighboring properties are affected by the same location characteristics, including a series of submarket dummy variables in a hedonic regression can capture location effects and reduce the number of explanatory variables. However, these submarket dummy variables are obtained from the administrative boundary, which is a predefined submarket. Bourassa, Hamelink, Hoesli and MacGregor (1999) and Tu, Sun and Yu (2007) noted that administrative boundary is not the optimal way to segment the market. This predefined boundary ignores spatial correlation among properties in different administrative segments, the properties located on the edge of one segment may have price correlation with its neighborhoods in the nearest segment. Pace, Barry and Sirmans (1998) argued that certain models that include a location indicator in the regression cannot, unfortunately, yield independent residuals. Many previous papers noted the problem with spatial correlation among neighboring properties, indicating that prices of nearby properties located in different administrative boundaries may be correlated. Tu, Sun and Yu (2007) confirmed this notion by showing that price predictions based on new market segmentation using geostatistic methodology are more precise that those based on administrative boundaries.

However, we follow the notion of Can (1992), who distinguished between two spatial correlation effects. The first one is the “neighborhood effect,” such as the impact

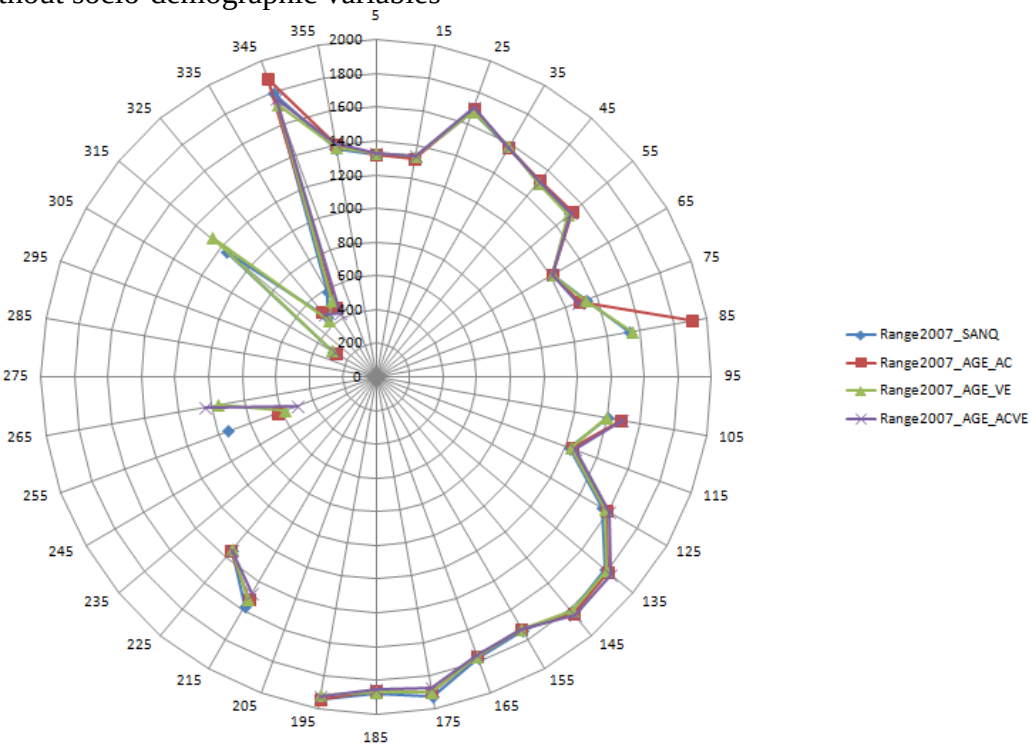
of shared neighborhood characteristics on housing prices. The second one, "adjacency effects," represents the absolute spatial spillover effects, i.e., the effect of prices of adjacent structures on the price of a given structure. These effects can cross neighborhood boundaries. We consider that arrondissement number included as an explanatory variable of the hedonic regression helps capture the correlation effect at the macro level or the neighborhood effects. Then, the correlation effect at the micro level or the adjacency effects is presented in the residuals of the hedonic regression. Moreover, this micro correlation effect may be stationary.

Comparing Figure III.11 to Figure III.13 (for the *Arc de Triomphe*) and Figure III.12 to Figure III.15 (for the *Place d'Italie*), when age of buyer, age of seller, or both, are included as an explanatory variable in a hedonic regression, the estimated range is not different from the results obtained from the standard hedonic model. Including socio-demographic variables in the hedonic regression cannot reduce the variability in the range. In addition, some windows still exist whose ranges are not estimable. Figure III.14 (for the *Arc de Triomphe*) and Figure III.16 (for the *Place d'Italie*) show the estimated ranges obtained when submarket variables are included in the hedonic regression for the *Arc de Triomphe* and the *Place d'Italie*. These two figures show that estimated ranges are significantly reduced when arrondissement number or district number is included as an explanatory variable. Table III.9 provides a statistical summary of the estimated ranges. For the segments around the *Arc de Triomphe* without any complementary variable, the 10th percentile of the estimated ranges is 424 meters and the 90th percentile[7] is 1,842 meters (with a mean of 1,413 meters and a standard deviation of 798 meters). The 10th percentile and the 90th percentile of the estimated ranges are reduced to 321 meters and 1,215 meters (with a mean of 865 meters and a standard deviation of 466 meters) when arrondissement dummies are included as explanatory variables in the hedonic model. These two percentiles are reduced to 218 meters and 885 meters if district dummies are included in the hedonic regression. For segments around the *Place d'Italie* without any complementary variable, the 10th percentile of the estimated ranges is 508 meters and 1,338 meters for the 90th percentile (with a mean of 1,006 meters and a standard deviation of 384 meters), and become 255 and 676 meters when arrondissement dummies

[7] The 10th and the 90th percentile of estimated ranges are chosen to exclude the extremum problem.

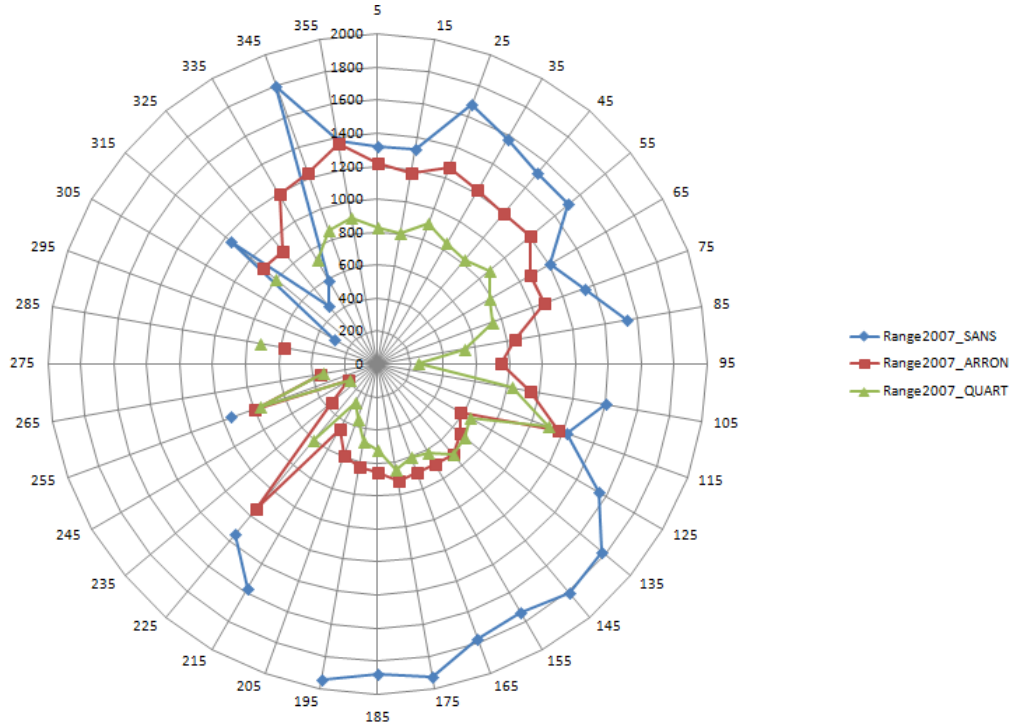
are included as explanatory variables. They are reduced to 214 meters and 653 meters when district dummies are included in the hedonic regression. Even when a spatial characteristic, such as arrondissement number or district number, is included in the regression, the degree of spatial correlation may be reduced but cannot be eliminated. The residuals still present a spatial correlation structure with a significant estimated range, but this distance limit is reduced. This reduction in the estimated range can be explained by the two levels of correlation effects.

Figure III.13: Polar chart of 36 estimated ranges around the *Arc de Triomphe*, 2007, with and without socio-demographic variables



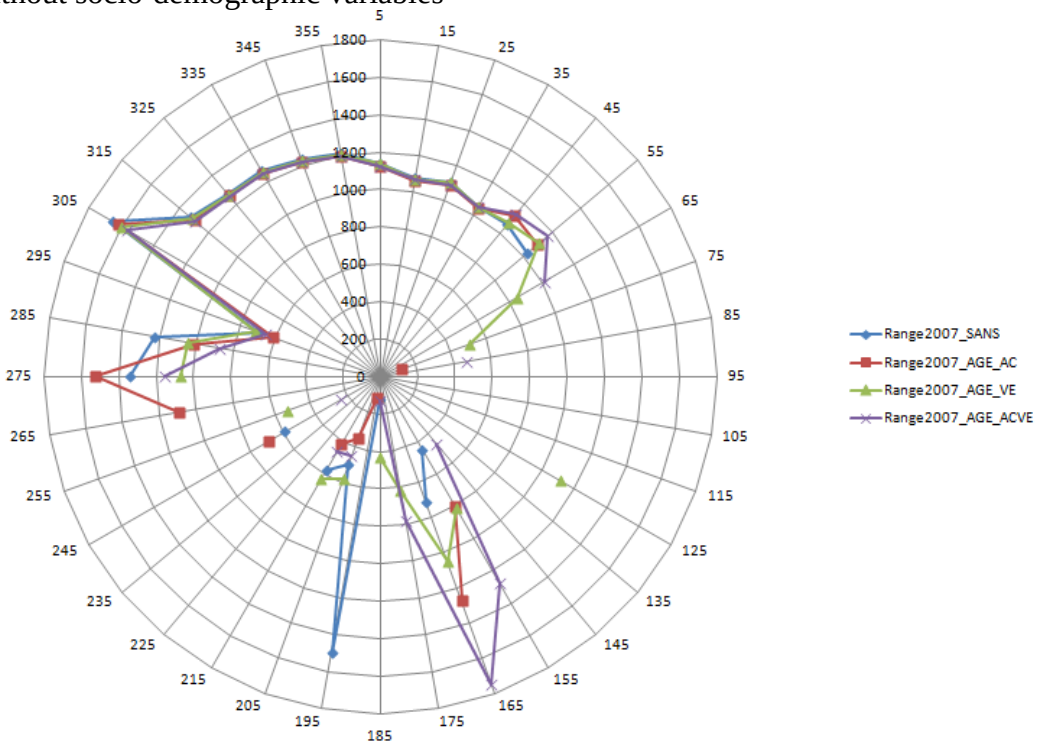
(Range2007_SANS: Standard model without socio-demographic variables; Range2007_AGE_AC: Includes age of buyer; Range2007_AGE_VE: Includes age of seller; Range2007_AGE_ACVE: Includes age of buyer and seller.)

Figure III.14: Polar chart of 36 estimated ranges around the *Arc de Triomphe*, 2007, with and without submarket variables



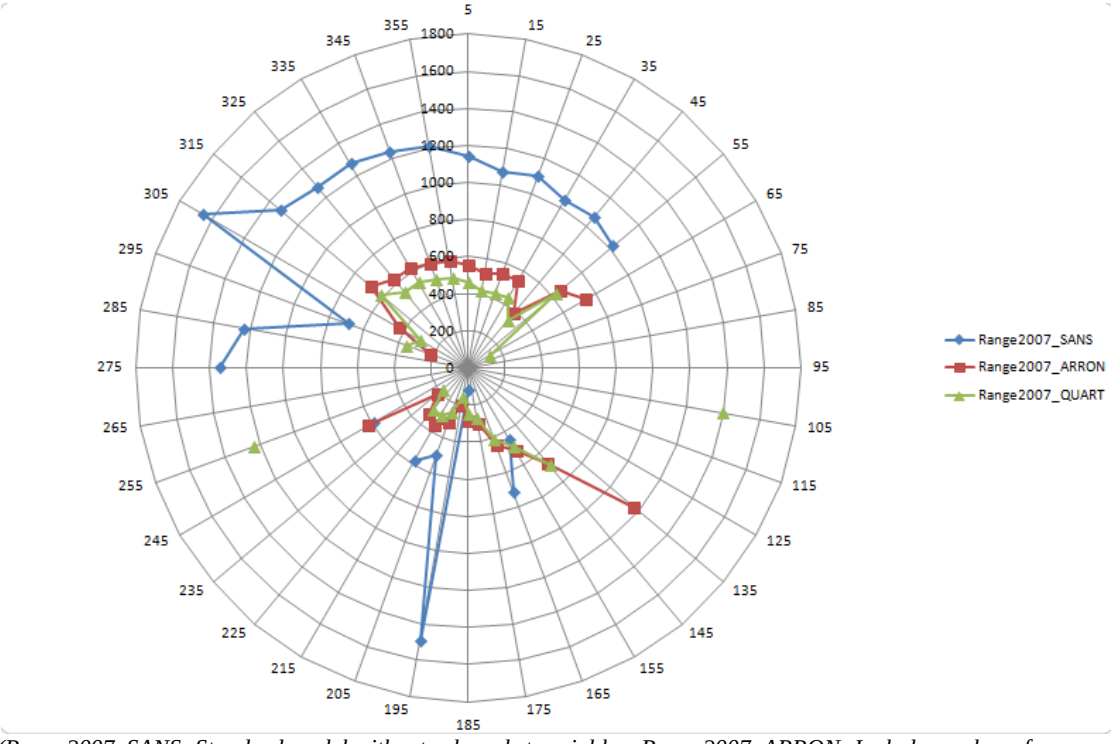
(Range2007_SphSANS: Standard model without submarket variables; Range2007_ARRON: Include number of arrondissement; Range2007_QUART: Include number of quarter.)

Figure III.15: Polar chart of 36 estimated ranges around the *Place d'Italie*, 2007, with and without socio-demographic variables



(Range2007_SANS: Standard model without socio-demographic variables; Range2007_AGE_AC: Include age of buyer; Range2007_AGE_VE: Include age of seller; Range2007_AGE_ACVE: Include age of buyer and seller.)

Figure III.16: Polar chart of 36 estimated ranges around the *Place d'Italie*, 2007, with and without submarket variables



(Range2007_SANS: Standard model without submarket variables; Range2007_ARRON: Include number of arrondissement; Range2007_QUART: Include number of quarter.)

Table III.9: Estimated ranges obtained from spherical and Gaussian fitted variograms around the *Arc de Triomphe* and the *Place d'Italie*, 2007

Arc de Triomphe	*SANS*	*AGE_AC*	*AGE_VE*	*AGE_ACVE*	*ARRON*	*QUART*
Minimum	301	281	310	291	200	197
Maximum	1940	1940	1921	1923	1355	1105
Mean	1413	1428	1381	1383	865	671
Standard Deviation	798	567	564	491	466	331
10th percentile	424	464	440	466	312	218
90th percentile	1842	1889	1831	1848	1215	885
Place d'Italie						
Minimum	119	119	433	127	206	132
Maximum	1657	1621	1605	1746	1170	1399
Mean	1006	974	974	966	505	465
Standard Deviation	384	409	299	408	209	290
10th percentile	508	367	548	454	255	214
90th percentile	1338	1293	1263	1284	676	653

(SANS: Standard model without socio-demographic or submarket variables; AGE_AC: Include age of buyer; AGE_VE: Include age of seller; AGE_ACVE: Include age of buyer and seller; ARRON: Include number of arrondissement; QUART: Include number of quarter; Sph: Spherical fitted variogram; and Gau: Gaussian fitted variogram.)

Regarding the spatial characteristic that causes correlations among property prices, this spatial effect should be separated in two groups: location characteristics that may cause price variability at the macro level and neighborhood characteristics that vary a property's price at the micro level. Submarket explanatory variables can capture a property's price corresponding to location characteristics but the spatial correlation remains in the residuals. Consequently, this spatial correlation among residuals is caused by neighborhood characteristics.

The first group, location characteristics, corresponds to the characteristics that affect property prices at the macro level, and property prices are correlated because they are located in the same area. Any indicator of a property's location (such as an indicator of administrative boundary, number of arrondissement or quarter, latitude and longitude), any indicator of accessibility (such as distance or travel time to CBD, to public transport, or to an employment center), and any area quality indicator (such as criminal rate, average income or education level) can be considered a location characteristic. Prices for properties with the same physical characteristics should differ in inner Paris and in outer Paris. Increasing taxes in an arrondissement only increases property prices within that

arrondissement. Including these location characteristics in a hedonic regression can measure the value of a spatial characteristic at the macro level; the spatial value remains in the residuals that are caused by the neighborhood characteristic. Prices of properties with same physical characteristics and that are located in the same arrondissement should be different for property near a natural park and property near a noisy factory. All apartments located in a luxury building will have higher prices than other apartments. Neighborhood characteristics may correspond to any indicator of neighbors, such as quality of the neighborhood building, building security, or cleanness of the streets. However, specifying and including all neighborhood characteristics in a hedonic regression is impossible. Moreover, each property has different neighborhood characteristics, which is why geostatistic analysis is applied in the second step to capture spatial correlation at the micro level caused by neighborhood characteristics or omitted variables.

7. Conclusion and others approaches

The stationary assumption is normally proposed because of its global homogeneity to an applied geostatistic approach. This stationary assumption can hold if the spatial data corresponds to the mineral resources. Different from mineral data, real estate data are heterogeneous and spatial factors affecting property prices may vary by location; thus, assuming a spatial stationarity assumption for real estate data seems difficult. This paper is probably a first attempt at examining the violation of the stationary assumption in space and time.

Our results show that the estimated range of a semivariogram varies over the studied period and over the region of interest. A stationary assumption cannot simply be assumed for real estate data and we should not compute a single common variogram for the entire studied period and for the entire region of interest. An adequate sample size should be applied to ensure local homogeneity and local stationarity. Others several possibilities for implementing a non-stationary variogram are proposed. Two common approaches include (i) segmentation and (ii) spatial transformation data.

Moreover, we attempt to identify the cause of variability in semivariograms. When arrondissement number or district number is included as an explanatory variable in a hedonic regression, the estimated range is reduced and becomes more homogeneous. We explain these results through the two levels of the correlation effect caused by spatial characteristics of property. The location characteristic may cause price variability at the macro level and the neighborhood characteristics may vary a property's price at the micro level. Submarket explanatory variables included in a hedonic regression can capture a property's price corresponding to location characteristics but the spatial correlation remains in the residuals. Consequently, this spatial correlation among residuals is caused by neighborhood characteristics.

CHAPITRE IV MODELE D'ECONOMETRIE SPATIALE ET ETUDE IMMOBILIERE

1. Introduction

Le modèle d'économétrie spatiale diffère de la géostatistique sur plusieurs points. La géostatistique est utilisée pour analyser des observations qui sont distribuées de façon aléatoire dans un espace continue, mais le modèle d'économétrie spatiale est mieux adapté aux données de type treillis, c'est-à-dire l'espace étudié correspond à des unités géographiques d'un réseau structuré. La géostatistique estime directement l'autocorrélation spatiale par l'analyse de variogramme. Inversement, l'économétrie spatiale détermine indirectement la dépendance spatiale en essayant de redéfinir les variables explicatives de l'équation de régression, afin de pouvoir capter le degré de corrélation spatiale. L'approche issue de l'économétrie spatiale est basée sur l'idée que l'autocorrélation spatiale se présente uniquement parmi des observations voisines, la prise en compte de la valeur des observations voisines dans le modèle de régression permettant de capter l'information liée à ce degré de dépendance. Il faut, par conséquent, ajouter dans le modèle de régression hédonique les éléments permettant de mesurer la relation entre les valeurs voisines. Ces éléments sont, premièrement, la matrice de poids spatiaux qui accordent une pondération différente à des observations voisines et, deuxièmement, le paramètre mesurant le degré de dépendance spatiale entre les observations voisines. Dans les faits, comparés à la géostatistique, l'économétrie spatiale est plus connue et plus utilisée dans l'étude de la dépendance spatiale des données immobilières. Soumis à des conditions moins restrictives que la géostatistique, l'économétrie spatiale ne nécessite pas de formuler des hypothèses sur le processus spatial. Cette méthode est plus simple et demande moins de temps de calcul. Néanmoins, la méthode de l'économétrie spatiale est basée sur certains choix arbitraires tels que le choix de la condition de voisinage, le choix de la pondération accordée à chaque observation voisine et le choix des caractéristiques présentant la dépendance spatiale.

L'objectif de ce chapitre est de présenter l'estimation du degré de corrélation spatiale avec la méthode de l'économétrie spatiale et de discuter sur le choix de modèle d'estimation. Ce chapitre est divisé en 4 sections. Après cette introduction, la seconde section présente les différents éléments de l'économétrie spatiale à savoir les deux processus spatiaux, le modèle de l'économétrie spatiale, la détermination de la matrice de voisinage et de la matrice de poids spatiaux et la méthode d'estimation. La troisième

section concerne l'application du modèle d'économétrie spatiale dans le cas de l'étude immobilière et les exemples de choix de modèle selon la source de la dépendance spatiale des prix immobiliers. La quatrième section correspond à la conclusion du chapitre.

2. Économétrie spatiale

Cette section a pour but de présenter l'approche issue de l'économétrie spatiale. Les deux processus spatiaux que sont le processus autorégressif et le processus moyen mobile sont présentés. Une fois spécifié le processus spatial approprié, il faut intégrer le processus choisi, soit dans la partie des variables endogènes, soit dans la partie des variables exogènes, soit dans la partie des erreurs, ou alors dans toutes les parties de l'équation de régression. Ce choix dépend, normalement, de la source de corrélation considérée. Nous présentons aussi les modèles de régression dans cette section. Ces modèles se différencient selon le processus spatial et les variables présentant la caractéristique de dépendance spatiale.

En plus de la description de chaque modèle, nous essayons aussi dans cette section de classer ces modèles selon les sources de l'autocorrélation spatiale. La matrice de poids spatiaux est un élément principal de cette méthode, elle est utilisée pour capter l'information liée au degré de corrélation spatiale. La détermination de cette matrice de poids dépend de la définition du voisinage et la définition de poids spatiaux. Nous terminerons cette section par le détail de la méthode d'estimation.

L'économétrie spatiale a été largement développée et il existe plusieurs modèles et méthodes d'estimation disponibles dans la littérature. Cette section ne présente cependant que les modèles ou les méthodes souvent utilisées en étude immobilière.

2.1. Processus spatiaux

L'étude de la dépendance spatiale commence par le choix de la structure spatiale entre les observations. Deux processus spatiaux qui sont souvent cités dans la littérature

(Anselin (1988); Dubin (1988); Gallo (2002)) sont le processus autorégressif spatial (*Spatial autorégressif* - SAR) et le processus moyenne mobile spatiale (*Spatial moving average* - SMA). Ces deux processus se différencient par la structure de corrélation. Le processus moyen mobile spatial s'interprète comme une interaction « locale » entre des voisines. Un choc sur une observation n'a d'impact que dans un voisinage de cette observation, par exemple dans le voisinage d'ordre 1 et 2 ; *le voisinage d'ordre 1* contient le voisin le plus proche et *le voisinage d'ordre 2* contient le voisin du voisin le plus proche. Par contre le processus autorégressif spatial (SAR) présente une interaction « globale ». Un choc sur n'importe quelle observation a une influence sur toutes les autres observations voisines mais cette influence se réduit en fonction de la distance séparant les observations. Une fois spécifié le processus spatial, ce processus peut être incorporé soit aux variables endogènes, soit aux variables endogènes, soit aux résidus du modèle de régression.

2.1.1. Processus autorégressif spatial (SAR)

Le *processus autorégressif spatial* proposé par Whittle (1954) est le processus le plus simple pour étudier la dépendance spatiale. Selon ce processus, la valeur d'une observation dépend des observations voisines. Ce processus considère que la valeur d'une observation a un influence pas seulement sur son voisin le plus proche mais que cette influence se diffuse aussi sur les observations voisines de son voisin le plus proche et ainsi de suite. L'importance de cette influence se réduit en fonction de la distance entre les observations. Ce processus est souvent utilisé dans l'étude de l'interaction « globale » entre les observations telle que la dépendance spatiale générale.

Le processus autorégressif spatial peut être défini par :

$$y_i = \rho \sum_{j=1}^{n} w_{ij} y_j + \varepsilon_i \qquad \text{Eq. IV.1}$$

$$\varepsilon_i \sim N(0, \sigma^2) \qquad i = 1, \dots, n$$

La valeur d'une observation y collectée au pointS_i, notée y_i est la somme des valeurs voisines pondérées, $\rho \sum_{j=1}^{n} w_{ij} y_j$. Le paramètre ρ est un indicateur du degré de dépendance spatiale qui peut avoir une valeur comprise entre 0 et 1. Une valeur élevée de ρ signifie que le degré de corrélation est très fort et donc l'observation y dépend fortement des observations voisines. w_{ij} est la pondération accordée à l'observation j dans le voisinage de l'observation i, il a une valeur supérieure à 0 si l'observation collectée au point S_j est dans le voisinage de l'observation collectée au pointS_i, et égale à 0 sinon. L'ensemble des pondérations w_{ij} permet de construire la matrice de poids spatiaux, notée W. Si le nombre d'observations est n, cette matrice W est de dimension $n \times n$ et elle contient les informations liés à la relation de dépendance entre tous les couples d'observations. Il existe plusieurs façons de déterminer les éléments de la matrice W, soit par la contiguïté, soit par la condition de distance. L'explication détaillée de la construction de matrice de poids spatiaux (W) est présentée dans la section 2.3 de ce chapitre.

Avec y le vecteur des observations de dimension $n \times 1$ et ε le vecteur des erreurs, de dimension $n \times 1$, le processus autorégressif spatial est décrit sous forme matricielle de façon suivant :

$$y = \rho W y + \varepsilon$$

alors

$$(I_n - \rho W) y = \varepsilon \qquad \text{Eq. IV.2}$$

L'Eq. IV.2 peut s'écrire

$$y = (I_n - \rho W)^{-1} \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2 I_n) \qquad \text{Eq. IV.3}$$

La matrice variance-covariance de y peut être donc écrite en fonction de deux paramètres, la variance des résidus ε et le coefficient de corrélation spatial (σ^2, ρ).

$$Cov(yy') = \sigma^2 \left((I_n - \rho W)' (I_n - \rho W) \right)^{-1} \qquad \text{Eq. IV.4}$$

En développant la matrice de variance-covariance de l'Eq. IV.4, nous pouvons mettre en exerce la structure de corrélation spatiale prise en compte par ce processus autorégressif spatial :

$$\begin{aligned}\left((I_n - \rho W)'(I_n - \rho W)\right)^{-1} &= I_n + \rho(W + W') \\ &+ \rho^2(WW + WW' + W'W') \\ &+ \rho^3(WWW + WWW' + WW'W') + \cdots\end{aligned} \qquad \text{Eq. IV.5}$$

À noter que la matrice W représente le voisinage d'ordre 1, W^2 représente donc le voisinage d'ordre 2 (les voisins d'un voisin) et ainsi de suite pour l'ordre supérieur, l'explication détaillée des ordres de voisinage se trouve dans la section 2.3.1. L'Eq. IV.5 illustre les deux propriétés de ce processus autorégressif. Premièrement, même si seulement la matrice de voisinage d'ordre 1 (W) s'affiche dans l'équation de régression (Eq. IV.3), ce processus prend bien en compte l'autocorrélation spatiale entre le voisinage d'ordre supérieur ($W^2, W^3, \ldots$). En effet, ce processus représente une interaction « globale » entre les observations. Deuxièmement, avec $|\rho| < 1$, le processus autorégressif spatial accorde un degré de corrélation très élevé au voisinage d'ordre 1 et l'ampleur de cette corrélation diminue pour le voisinage d'ordre supérieur ($|\rho| > |\rho|^2 > |\rho|^3 \cdots$). Ce processus autorégressif est souvent utilisé pour expliquer la corrélation causée par la diffusion spatiale (*spillover effect*) ce qui se traduit ici par le fait qu'une observation a une influence très forte sur son proche voisin et que cette influence se diffuse dans le voisinage lointain.

En étude immobilière, le processus autorégressif est souvent utilisé dans le cas où l'autocorrélation spatiale est causée par un processus d'évaluation tel que le propriétaire ou l'expert immobilier estime la valeur d'un bien immobilier par référence à des valeurs de biens voisins. Un exemple est le cas de l'autocorrélation de prix causée par l'existence d'un centre commercial. Le prix des biens immobiliers situés très proche de ce centre sont fortement corrélés grâce à la facilité d'accès au service et cette corrélation se diffuse même sur les autres biens immobiliers situés plus loin, mais le degré de corrélation diminue avec la distance par rapport à la localisation de centre commercial.

Le processus autorégressif peut être incorporé soit dans la partie des variables endogène (modèle SAR présenté dans la section 2.2.1) ou dans la partie des erreurs du modèle (modèle SEM présenté dans la section 2.2.3) ou dans ces deux parties (modèle GSM présenté dans la section 2.2.5). Ci-dessous les trois modèles cités.

Modèle SAR : $y = \alpha\iota_n + \rho Wy + X\beta + \varepsilon,\ \varepsilon \sim N(0, \sigma^2 I_n)$

Modèle SEM : $\varepsilon = \theta M\varepsilon + \epsilon, \epsilon \sim N(0, \sigma^2 I_n)$

Modèle GSM : $y = \alpha\iota_n + \rho Wy + X\beta + \varepsilon$ et $\varepsilon = \theta M\varepsilon + \epsilon, \epsilon \sim N(0, \sigma^2 I_n)$

Les matrices W et M sont les matrices de poids. Les paramètres ρ et θ sont les coefficients du degré de corrélation spatiale. α est la constante du modèle. X est la matrice des variables explicatives, de dimension $n \times k$ et β est le vecteur des coefficients de dimension $k \times 1$. ι_n est le vecteur identité de dimension $n \times 1$. La matrice I_n représente la matrice identité de dimension $n \times n$.

2.1.2. Processus moyenne mobile spatiale (SMA)

Un autre processus souvent utilisé dans l'étude de l'autocorrélation spatiale des prix immobiliers est le processus moyenne mobile spatiale (SMA) proposé par Haining (1993). Ce processus est plus approprié dans le cas de l'étude de l'autocorrélation causé par un choc ou par un effet lié à une externalité. Cet effet d'externalité se présente donc dans la partie des erreurs de modèle.

$$\begin{aligned} \varepsilon &= \rho W\epsilon + \epsilon \\ \epsilon &\sim N(0, \sigma^2 I_n) \end{aligned} \qquad \text{Eq. IV.6}$$

ρ est le degré de corrélation spatiale, ε et ϵ sont les vecteurs des erreurs de dimension$n \times 1$, et W la matrice de voisinage, de dimension $n \times n$. La matrice de variance-covariance des erreurs de l'estimation est:

$$Cov(\varepsilon\varepsilon') = \sigma^2(I_n + \rho(W + W') + \rho^2 WW') \qquad \text{Eq. IV.7}$$

La matrice de variance-covariance de l'Eq. IV.7 montre bien la différence entre les processus SAR et SMA. Le processus moyenne mobile ne prend uniquement en compte que l'autocorrélation entre le voisinage d'ordre 1 (W) et d'ordre 2 (WW'). Comparé au processus autorégressif qui étudie l'interaction global, ce processus moyen mobile est plus approprié dans le cas d'une étude de l'interaction « local » entre les observations. La valeur d'une observation a une influence uniquement sur le voisinage de cette

observation. Le processus moyen mobile peut être appliqué dans une étude immobilière par exemple pour mesurer l'autocorrélation causé par une externalité négative, telle que la nuisance sonore ou la pollution dans le cas où ces biens sont situés près d'une usine.

Nous remarquons que, comme dans l'analyse des séries temporelles, il est possible de décrire le processus autorégressif spatial comme une combinaison des processus moyenne mobile. L'ensemble des interactions locales donne l'interaction globale entre les observations.

$$y = (I_n - \rho W)^{-1}\varepsilon = \varepsilon + \rho W\varepsilon + \rho^2 W^2\varepsilon + \cdots \qquad \text{Eq. IV.8}$$

2.2. Modèles d'économétrie spatiale

Étant donné que les sources de la dépendance spatiale sont variées, les statisticiens ont développé différents modèles d'économétrie spatiale permettant de prendre en compte, dans la régression linéaire, la dépendance selon la source de l'autocorrélation spatiale. Le processus spatial décrit dans la section précédente peut être alors intégré au modèle de régression de plusieurs façons ; soit dans la partie variable à expliquer, soit dans la partie des variables explicatives, soit dans la partie des résidus de l'estimation. Les modèles spatiaux diffèrent l'un de l'autre en fonction de deux conditions. Premièrement, la partie qui contient l'autocorrélation spatiale et, deuxièmement, le type de processus spatial intégré. Plusieurs modèles sont mentionnés dans la littérature. Gallo (2002) décrit le modèle[8] autorégressif spatial (SAR) et le modèle d'erreurs spatial (SEM). Anselin (1988) introduit le modèle SAR, le modèle autorégressif et moyenne mobile spatial (SARMA) et Kelejian et Robinson (1993) proposent le modèle de variables exogènes décalées (SLX). LeSage et Pace (2009) présentent plusieurs modèles, parmi lesquels le modèle de Durbin spatial (SDM), le modèle de Durbin et d'erreurs spatial (SDEM) et le modèle spatial général (GSM). Ces modèles sont brièvement détaillés dans la suite.

[8] SAR est l'acronyme pour *Spatial Autoregressive Model*, SEM est l'acronyme pour *Spatial Error Model*, SARMA est l'acronyme pour *Spatial Autoregressive Moving Average Model*, SLX est l'acronyme pour *Spatial Lagged X Model*, SDM est l'acronyme pour *Spatial Durbin Model*, SDEM est l'acronyme pour *Spatial Durbin Error Model* et GSM est l'acronyme pour *General Spatial Model*.

Une des sources de l'autocorrélation spatiale la plus observée dans le cas de l'étude immobilière est le processus d'évaluation. La valeur de prix immobilier dépend du prix des biens voisins car pour déterminer le prix de vente de son bien, le propriétaire ou l'expert se renseigne souvent auprès des voisins. Le modèle d'économétrie spatiale permettant d'analyser ce problème est le modèle autorégressif spatial (SAR) qui intègre le processus spatial dans la partie des variables endogènes, présenté dans la section 2.2.1.

La ressemblance des caractéristiques de biens voisins est une autre source de la dépendance spatiale ; l'autocorrélation spatiale se présente donc dans la partie des variables explicatives. En étude immobilière, la qualité du quartier est une des caractéristiques qui a une influence sur le prix immobilier mais ce n'est pas uniquement la qualité du quartier où le bien est localisé, la qualité des quartiers voisins peut avoir une influence sur le prix. Le modèle qui permet d'analyser la dépendance des variables explicatives est le modèle de variables exogènes décalées (SLX) présenté dans la section 2.2.2.

La troisième source de la dépendance spatiale est certaines variables omises lors de la définition du modèle de régression étant donné qu'il est totalement impossible de pouvoir ajouter toutes les variables explicatives nécessaires dans l'équation. Par exemple, l'existence proche d'un accès aux transports en commun peut créer une autocorrélation de prix parmi les biens localisés aux environs de cet accès mais si cette information n'est pas disponible dans la base de données, l'absence de cette variable explicative peut créer le problème de dépendance spatiale des résidus de l'estimation. Le modèle de l'économétrie spatiale développé pour analyser ce problème est le modèle d'erreurs spatiale (SEM) présenté dans la section 2.2.3.

Une combinaison de ces différents modèles est possible. Par exemple, le modèle de Durbin spatial (SDM) qui présente à la fois la dépendance spatiale des variables explicatives et des variables endogènes est présenté dans la section 2.2.4. Les modèles qui présentent l'autocorrélation spatiale des variables endogènes et des résidus sont le modèle spatial général (GSM) et le modèle autorégressif et moyenne mobile spatial (SARMA) présentés dans les sections 2.2.5 et 2.2.6. Le modèle de Durbin et d'erreurs spatial (SDEM) détaillé dans la section 2.2.7 prend en compte l'autocorrélation spatiale des variables explicatives et des résidus.

2.2.1. Modèle autorégressif spatial (SAR)

Les valeurs immobilières des biens voisins sont souvent corrélées parce que le propriétaire ou l'expert immobilier estime la valeur d'un bien en prenant comme référence la valeur des biens voisins. Dans l'équation de régression entre le prix immobilier et ses caractéristiques, la variable dépendante contient donc la dépendance spatiale.

Le prix immobilier y ne dépend pas uniquement des caractéristiques physiques données dans la matrice X mais elle dépend aussi des observations voisines. Le processus spatial est ajouté dans la partie variable dépendante. Le modèle autorégressif spatial (SAR) se présente donc comme le modèle de régression linéaire simple tout en intégrant le terme $\rho W y$ qui contient les poids des observations voisines intégrées dans l'estimation. Ce modèle s'écrit donc l'Eq. IV.9:

$$y = \alpha \iota_n + \rho W y + X\beta + \varepsilon \qquad \text{Eq. IV.9}$$

L'Eq. IV.9 peut être développée comme :

$$y = (I_n - \rho W)^{-1}(\alpha \iota_n + X\beta) + (I_n - \rho W)^{-1}\varepsilon \qquad \text{Eq. IV.10}$$

avec $\varepsilon \sim N(0, \sigma^2 I_n)$

Comme dans la régression hédonique standard, y est le vecteur $n \times 1$ contenant l'ensemble des prix immobiliers, X la matrice des caractéristiques des biens de dimension $n \times k$, $\iota_n$est le vecteur unité de dimension $n \times 1$ qui ne contient que la valeur 1 à chaque ligne, $I_n$est la matrice identité de dimension $n \times n$. Les paramètres à estimer sont les coefficients de la régression hédonique standard α et β. Le vecteur ε de dimension $n \times 1$ contient les résidus de l'estimation, ces résidus sont supposés indépendants et identiquement distribués, $\varepsilon \sim N(0, \sigma^2 I_n)$. La matrice W est la matrice de poids spatiaux. L'élément $w_{i,j}$ de cette matrice indique la façon dont l'observationy_i (collectée au S_i) est spatialement corrélée à l'observation y_j (collectée au S_j). Un coefficient additionnel est le paramètre ρ qui mesure le degré de dépendance spatiale des observations, $|\rho| < 1$. Si ce coefficient ρ n'est pas significativement différent de 0, il y a pas corrélation spatiale entre les prix immobiliers, dans ce cas le résultat de la régression obtenu est le même que celui de MCO. Le vecteur ε correspond aux résidus de l'estimation.

La dépendance entre les variables endogènes est le cas le plus souvent analysé et le modèle SAR est le plus utilisé en étude immobilière. Cependant il existe d'autres modèles qui se différencient par la source de dépendance spatiale.

2.2.2. Modèle de variables exogènes décalées (SLX)

Les caractéristiques de biens voisins sont habituellement similaires dans la mesure où ces biens sont construits dans la même période. Ils ont donc le même âge et la même structure. Une autre situation possible est l'autocorrélation causée par une externalité, positive ou négative, qui affecte le prix des biens dans une zone. Le processus spatial est donc ajouté dans la partie des variables explicatives de l'équation de régression pour prendre en compte l'autocorrélation créée par cette similitude. En notant β_1 les coefficients correspondant aux variables explicatives ne présentant pas la dépendance spatiale, et β_2 les coefficients correspondant aux variables explicatives présentant l'autocorrélation spatiale, nous pouvons réécrire l'équation de régression de la façon suivante :

$$y = \alpha\iota_n + X\beta_1 + WX\beta_2 + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2 I_n) \qquad \text{Eq. IV.11}$$

LeSage et Pace (2009) appellent ce modèle, le modèle de variables exogènes décalées (*spatial lag X model*) ou SLX.

2.2.3. Modèle d'erreurs spatiales (SEM)

Une autre difficulté dans la définition du modèle de régression est de déterminer le nombre de variables explicatives à ajouter dans le modèle. Un modèle avec très peu de variables explicatives ne produit pas de résultat fiable. Par contre, un nombre trop important de variables explicatives peut créer un problème de multi-colinéarité et peut aussi réduire le pouvoir explicatif du modèle. Selon LeSage et Pace (2009), le modèle de régression par les MCO ne peut pas prendre en compte, dans la régression, toutes les caractéristiques spécifiques des observations. D'après cet article, certaines variables

explicatives qui présentent le caractère de dépendance spatiale peuvent être ignorées lors de la définition du modèle à cause d'une mauvaise définition du modèle ou d'un manque d'information. Dans ce cas, la dépendance spatiale se présente donc dans la partie des résidus de l'estimation.

Supposons que la valeur d'un bien dépend de deux types de caractéristiques, par exemple, les caractéristiques physiques, données par la matrice *X*, et les caractéristiques spatiales données par la matrice *Z*. La valeur de y est définie par :

$$y = X\beta + Z\gamma \qquad \text{Eq. IV.12}$$

Remarquons que *X* et *Z* peuvent être indépendantes ou dépendantes, ces deux situations sont envisageables et le développement du modèle se diffère d'un cas par rapport à l'autre.

Si la base de données existante manque d'information sur les caractéristiques spatiales (la matrice *Z*), la valeur correspondante aux caractéristiques de localisation se présente donc dans les erreurs de l'estimation et le modèle devient donc :

$$y = \beta X + \varepsilon \qquad \text{Eq. IV.13}$$

Plusieurs situations sont envisageables. La première situation la plus simple est que les matrices *X* et *Z* soient indépendantes et qu'elles ne présentent pas de structure spatiale. Les erreurs de l'estimation ε sont indépendantes et identiquement distribuées, $\varepsilon \sim N(0, \sigma^2 I_n)$. L'estimateur MCO issu de l'Eq. IV.13 n'est pas biaisé mais il est possible que le pouvoir explicatif du modèle soit faible car s'il manque certaines variables explicatives nécessaires.

Le deuxième scénario est que les matrices *X* et *Z* sont indépendantes mais la matrice *Z* présente la structure spatiale. Cette situation est courante dans le cas d'une étude immobilière ; par exemple, si *X* représente l'ensemble des caractéristiques physiques du bien immobilier et que ces caractéristiques physiques ne sont pas spatialement corrélées. Par contre, la matrice *Z* qui représente l'ensemble des caractéristiques spatiales peut avoir la structure de dépendance spatiale. Si, par exemple, un élément de la matrice *Z* est la qualité du quartier, il est envisageable que la qualité d'un quartier dépende aussi de celle des quartiers voisins. Autre exemple, la fluidité du trafic

en ville a une influence non seulement sur les prix de biens immobiliers localisés dans cette ville, mais aussi sur le prix des biens localisés dans les villes voisines car la fluidité du trafic dans une ville permet de mesurer la rapidité d'accès aux villes voisines. Nous pouvons donc définir le processus autorégressif spatial des caractéristiques de localisation (la matrice *Z*) de la façon suivante :

$$Z = \rho WZ + \upsilon \qquad \text{Eq. IV.14}$$

Le processus spatial de Z est défini par :

$$Z = (I_n - \rho W)^{-1}\upsilon$$
$$\upsilon \sim N(0, \sigma^2 I_n) \qquad \text{Eq. IV.15}$$

La matrice *W* permet cette fois ci d'intégrer la dépendance spatiale des éléments de la matrice *Z*. Si les caractéristiques de la matrice Z sont ignorées lors de la définition du modèle et elles présentent la dépendance spatiale alors les erreurs de l'estimation ε contiennent l'autocorrélation spatiale. Le modèle d'erreurs spatial (SEM) se présente par :

$$y = \alpha\iota_n + X\beta + \varepsilon$$
$$\varepsilon = \rho W\varepsilon + \epsilon$$
$$\epsilon \sim N(0, \sigma^2 I_n) \qquad \text{Eq. IV.16}$$

Rappelons que l'estimateur par les MCO est non biaisé mais il n'est pas efficient et donc les résidus de l'estimation sont spatialement corrélés. Le recours à l'estimation avec moindres carrés généralisés est donc nécessaire. L'Eq. IV.16 peut être développée de la façon suivante afin d'obtenir les résidus de l'estimation indépendants :

$$\varepsilon = (I_n - \rho W)^{-1}\epsilon$$
$$y = \alpha\iota_n + X\beta + (I_n - \rho W)^{-1}\epsilon \qquad \text{Eq. IV.17}$$
$$\epsilon \sim N(0, \sigma^2 I_n)$$

2.2.4. Modèle de Durbin spatial (SDM)

À partir du modèle de l'autocorrélation spatiale des erreurs (SEM) présenté dans l'Eq. IV.16, nous pouvons envisager la situation où les résidus ε dépendent de certaines variables explicatives. Reprenons l'exemple de la qualité du quartier dans l'étude

immobilière. Si l'information concernant la qualité du quartier est manquante lors de la définition du modèle, les erreurs de l'estimation présentent alors le problème d'autocorrélation spatiale. En plus, cette qualité du voisinage peut dépendre du taux de criminalité de ce quartier, de la propreté du quartier, du revenu moyen des habitantes, etc. Supposons que ces informations soient disponibles et donc peuvent être pris en compte dans la définition du modèle. Les erreurs de l'estimation dépendent des variables explicatives du modèle. Le terme $X\gamma$ est ajouté dans l'équation des erreurs de l'Eq. IV.16.

$$\varepsilon = \rho W\varepsilon + X\gamma + \epsilon$$
$$\varepsilon = (I_n - \rho W)^{-1}(X\gamma + \epsilon) \qquad \text{Eq. IV.18}$$

En remplaçant l'Eq. IV.18 dans le modèle de régression standard, nous obtenons :

$$y = \alpha\iota_n + X\beta + (I_n - \rho W)^{-1}(X\gamma + \epsilon) \qquad \text{Eq. IV.19}$$

Le système d'équations du modèle de Durbin spatial est alors :

$$y = \alpha\iota_n + \rho Wy + X(\beta + \gamma) + \rho WX\beta + \epsilon$$
$$\epsilon \sim N(0, \sigma^2 I_n) \qquad \text{Eq. IV.20}$$

L'Eq. IV.20 inclut à la fois la dépendance spatiale des variables endogènes (les variables à expliquer, vecteur *y*) ainsi que celle des variables exogènes (variables explicatives, matrice *X*). Ce modèle paraît comme la combinaison du modèle SAR et du modèle SLX.

Généralement, le modèle de Durbin spatial (SDM) est défini de la façon suivante :

$$y = \alpha\iota_n + \rho Wy + X\beta + \theta WX + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2 I_n) \qquad \text{Eq. IV.21}$$

Pour résumer, dans le cas où certaines variables spatiales sont omises lors de la spécification du modèle, il y a trois scénarii possibles. Dans le premier scénario, si ces variables omises sont indépendantes des variables explicatives et si elles ne présentent pas de dépendance spatiale, alors le modèle de régression hédonique standard permet d'obtenir des estimations sans biais, mais à faible pouvoir d'explicatif. Dans le deuxième scénario, si les variables omises sont indépendantes des variables explicatives mais présentent le caractère de dépendance spatiale, les erreurs de l'estimation sont alors

corrélées et le modèle SEM est développé pour prendre en compte cette autocorrélation spatiale. Le troisième scénario est le cas où les variables omises sont dépendantes des variables explicatives et présentent le problème d'autocorrélation spatiale ; c'est le cas du modèle SDM.

Remarquons que le modèle de variables exogènes décalées (SLX présenté dans le point 2.2.2) correspond au modèle de Durbin spatial (SDM) qui exclut la dépendance des résidus. En effet, l'Eq. IV.11 correspond à l'Eq. IV.21 avec $\theta = 0$.

2.2.5. Modèle combiné – Modèle spatial général (GSM)

Considérons deux modèles présentés précédemment. Le modèle SAR qui prend en compte la dépendance spatiale des variables endogène, et le modèle SEM qui prend en compte la dépendance spatiale des erreurs du modèle. Nous pouvons envisager la combinaison de ces deux modèles de sorte que la dépendance spatiale se présente à la fois la partie des variables endogènes et la partie des erreurs de l'estimation. Reprenons les modèles SAR et SEM de l'Eq. IV.10 et l'Eq. IV.17 :

SAR : $y_{SAR} = (I_n - \rho W)^{-1}(\alpha \iota_n + X\beta) + (I_n - \rho W)^{-1}\varepsilon$

SEM : $y_{SEM} = \alpha \iota_n + X\beta + (I_n - \rho W)^{-1}\varepsilon$

LeSage et Pace (2009) proposent une approche Bayésienne qui combine ces deux modèles. Cette approche permet de calculer les probabilités à postériori, π_{SAR} et π_{SEM}, représentant les poids de chacun des modèles SAR et SEM dans ce modèle combiné. Le modèle combiné peut donc être exprimé à partir de deux équations (Eq. IV.10 et Eq. IV.17) comme :

$$y = \pi_{SAR} y_{SAR} + \pi_{SEM} y_{SEM} \qquad \text{Eq. IV.22}$$

Ce qui donne

$$y = R^{-1}\pi_{SAR}(\alpha \iota_n + X\beta) + \pi_{SEM}(\alpha \iota_n + X\beta) + R^{-1}(\pi_{SAR} + \pi_{SEM})\varepsilon \quad \text{avec } R = I_n - \rho W \qquad \text{Eq. IV.23}$$

En supposant que le modèle proposé ne peut être combiné que par ces deux modèles, nous obtenons une condition supplémentaire que $\pi_{SAR} + \pi_{SEM} = 1$. L'Eq. IV.23 peut donc être simplifiée :

$$\begin{gathered} Ry = \pi_{SAR}(\alpha\iota_n + X\beta) + R\pi_{SEM}(\alpha\iota_n + X\beta) + \varepsilon \\ (I_n - \rho W)\mathrm{y} = \pi_{SAR}(\alpha\iota_n + X\beta) + (I_n - \rho W)\pi_{SEM}(\alpha\iota_n + X\beta) + \varepsilon \\ (I_n - \rho W)\mathrm{y} = \alpha\iota_n + X\beta + \rho\pi_{SEM}WX\beta + \varepsilon \end{gathered} \qquad \text{Eq. IV.24}$$

Le système d'équations du modèle spatial général est le suivant :

$$\begin{gathered} \mathrm{y} = \alpha\iota_n + \rho Wy + X\beta_1 + WX\beta_2 + \varepsilon \\ \varepsilon \sim N(0, \sigma^2 I_n) \end{gathered} \qquad \text{Eq. IV.25}$$

avec $\pi_{SAR} + \pi_{SEM} = 1$ et $\rho\pi_{SEM}\beta = \beta_2$

Nous remarquons que les matrices de poids (W) des modèles SAR et SEM sont supposées égales dans l'Eq. IV.29. Il est cependant possible de supposer que ces deux matrices de poids sont différentes. Soient alors W_{SAR} et W_{SEM}, les matrices de poids spatiaux correspondants aux modèles SAR et SEM, ρ et θ, les deux coefficients du degré de dépendance spatiale. L'équation de l'estimation devient alors :

$$\begin{gathered} y = (I_n - \rho W_{SAR})^{-1}(\alpha\iota_n + X\beta) + (I_n - \rho W_{SAR})^{-1}(I_n - \theta W_{SEM})^{-1}\varepsilon \\ \varepsilon \sim N(0, \sigma^2 I_n) \end{gathered} \qquad \text{Eq. IV.26}$$

Le système d'équations présenté dans Eq. IV.26 permet de prendre en compte à la fois la dépendance spatiale des variables endogènes et celle des erreurs de l'estimation. Elle est considérée comme le modèle général de la dépendance spatiale (GSM).

2.2.6. Modèle autorégressif et moyenne mobile spatial (SARMA)

Le modèle SARMA ou *Spatial Autoregressive Moving Average* est un autre modèle combiné qui permet de prendre en compte à la fois la dépendance des variables endogènes et celle des erreurs de l'estimation comme le modèle GSM. Une seule différence apparaît dans la dépendance des erreurs, le modèle GSM intégrant le processus autorégressif dans les résidus. Inversement, le modèle SARMA intègre le processus

moyen mobile dans les résidus. Le système d'équations de SARMA est décrit de façon suivante :

$$\begin{aligned} y &= \alpha\iota_n + \rho W y + \varepsilon \\ \varepsilon &= (I_n - \rho\theta W)\epsilon \end{aligned} \qquad \text{Eq. IV.27}$$

En développant l'Eq. IV.27, le modèle SARMA est représenté par le système d'équations suivant :

$$\begin{aligned} y = (I_n - \rho W)^{-1}(\alpha\iota_n + X\beta) + (I_n - \rho W)^{-1}(I_n - \theta W)\epsilon \\ \epsilon \sim N(0, \sigma^2 I_n) \end{aligned} \qquad \text{Eq. IV.28}$$

Nous remarquons bien que pour le processus autorégressif, le facteur d'actualisation des observations voisines est égal à $(I_n - \rho W)^{-1} = I_n + \rho W + \rho W^2 + \rho W^3 + \cdots$. Pour le processus moyen mobile spatial, ce facteur est $I_n - \theta W$. Par conséquent, le processus autorégressif s'intéresse à la dépendance entre les voisins de second degré et plus, mais le processus moyenne mobile ne prend en compte que la dépendance des voisins de premier degré. C'est la raison pour laquelle, Anselin (1988) indique que le processus moyenne mobile permet de capter l'effet local causé par l'apparition immédiate d'une observation voisine, par contre le processus autorégressif va capter l'effet global.

2.2.7. Modèle de Durbin et d'erreurs spatiales (SDEM)

Dans la section 2.2.4, nous présentons le modèle de Durbin spatial (SDM) où la dépendance se trouve parmi les variables exogènes et les variables endogènes. LeSage et Pace (2009) développent le modèle de Durbin et d'erreurs spatiales (SDEM) qui intègre l'autocorrélation spatiale des variables exogènes et celle des erreurs de l'estimation. Le système d'équation Eq. IV.29 représentant les modèles SLX et SEM est :

$$\begin{aligned} y &= \alpha\iota_n + X\beta + WX\gamma + \varepsilon \\ \varepsilon &= (I_n - \rho W)^{-1}\epsilon \end{aligned} \qquad \text{Eq. IV.29}$$

En développant le système d'équations Eq. IV.29, on obtient le modèle de Durbin et d'erreurs spatial (SDEM) :

$$y = \alpha\iota_n + X\beta + WX\gamma + (I_n - \rho W)^{-1}\epsilon$$
$$\epsilon \sim N(0, \sigma^2 I_n) \qquad \text{Eq. IV.30}$$

Le Tableau IV.1 présenté dans la section 3.2 regroupe ces différents modèles d'économétrie spatiale disponibles dans la littérature.

2.3. Détermination de la matrice de poids spatiaux (W)

Pour tous les modèles présentés dans la section précédente, la matrice de poids spatiaux (*Spatial Weight Matrix*) est un élément important dans l'étude de dépendance spatiale. Cette matrice permet de prendre en compte l'interdépendance spatiale. La matrice de poids spatiaux notée *W*, est de dimension $n \times n$ où *n* le nombre d'observations considérées. Les éléments de cette matrice vérifient $w_{ij} > 0$ si les observations j et i respectent la condition d'interdépendance fixée au préalable, et $w_{ij} = 0$ sinon. Cette matrice a une taille très élevée car elle contient les informations sur chaque couple d'observations. En effet, cette matrice doit prendre en compte $n^2 - n$ relations possibles. Ce nombre élevé ($n^2 - n$) de relations de dépendance est difficilement définissable pour une base de données de grande taille. Ces relations de dépendance peuvent être réduites à $\frac{n^2-n}{2}$ si la condition de dépendance est symétrique. Une autre possibilité serait de mettre des contraintes sur les éléments de cette matrice de poids spatiaux. Plusieurs conditions d'interdépendance ont été proposées et elles dépendent de la position relative des observations. Par exemple les éléments de la matrice prennent une valeur 1, $w_{ij} = 1$, si deux observations se localisent dans le même quartier ou si deux observations se séparent moins de 500 mètres et $w_{ij} = 0$ sinon. Cette condition de voisinage permet de réduire le nombre de relations à déterminer. La matrice de poids spatiaux est normalisée de façon à obtenir la somme de chaque ligne égale à 1. Par conséquent, chaque élément de la matrice, w_{ij}, sera divisé par la somme totale de sa ligne $\sum_{j=1}^{n} w_{i,j}$.

La détermination des éléments de la matrice de poids peut se décrire en deux étapes : la détermination de l'ensemble des voisinages et la détermination des poids associés à chaque voisinage. La première étape consiste à déterminer, pour chaque observation, quel est son voisinage. Le voisin d'une observation peut être défini comme

l'observation la plus proche (le voisinage d'ordre 1), le voisin le plus proche du voisin d'ordre 1 (le voisinage d'ordre 2), le voisin le plus proche du voisin d'ordre *k* (le voisinage d'ordre *k+1*) ou les voisins qui se trouvent dans un rayon de *x* mètres autour de cette observation. L'ensemble des voisinages permet de créer la matrice de voisinage qui est une matrice binaire : les éléments de la matrice *B* vérifient $b_{i,j} = 1$ si l'observation *j* est un voisin de l'observation *i* et $b_{i,j} = 0$ sinon. La deuxième étape consiste à attribuer un poids à chaque voisin. La matrice de poids (*W*) contient les pondérations associées à tous les coups de voisins qui peuvent être égales pour tout couple de voisins ou variable suivant la distance séparant le couple.

2.3.1. Matrice de voisinage

Dans la littérature, deux principales conditions sont souvent utilisées pour déterminer l'ensemble des voisins à savoir la contiguïté et la condition de distance. Le choix entre ces deux conditions dépend de l'objectif de l'étude, du type de données spatiales et des informations disponibles dans la base de données.

La contiguïté

La contiguïté est la condition la plus simple à utiliser pour déterminer l'ensemble des voisinages. Elle ne nécessite pas d'information précise sur la localisation des observations. Cette condition est habituellement appliquée aux données spatiales agrégées ou aux données distribuées de façon polygonale. Les éléments de la matrice de contiguïté prennent une valeur 1 si deux observations sont les voisins et 0 sinon. Si la base de données contient, par exemple, une observation pour chaque région, les éléments de la matrice de contiguïté prend une valeur 1 si deux régions ont une frontière commune et 0 sinon.

Il existe plusieurs façons de définir la contiguïté. Certains exemples dans la littérature sont :

- *Linear contiguity* : les éléments de la matrice de contiguïté vérifient $w_{i,j} = 1$ si les observations *i* et *j* ont au moins une frontière commune, à gauche ou à droite, et $w_{i,j} = 0$ sinon,

- *Rock contiguity* : les éléments de la matrice de contiguïté vérifient $w_{i,j} = 1$ si les observations i et j ont au moins une frontière commune et $w_{i,j} = 0$ sinon,
- *Bishop contiguity* : les éléments de la matrice de contiguïté vérifient $w_{i,j} = 1$ les observations i et j ont au moins un coin commun et $w_{i,j} = 0$ sinon,
- *Queen contiguity* : les éléments de la matrice de contiguïté vérifient $w_{i,j} = 1$ les observations i et j ont au moins une frontière commune ou un coin commun et $w_{i,j} = 0$ sinon.

Remarquons qu'une région ne pouvant être contiguë avec elle même, on a $w_{i,i} = 0 \; \forall i$, et les éléments diagonaux de la matrice W sont nuls.

Les définitions de la contiguïté ci-dessus correspondent en fait à la notion de contiguïté d'ordre 1. Cette définition peut être généralisée à la notion de contiguïté d'ordre supérieur. Chez Gallo (2002), deux régions i et j sont contiguës d'ordre k si k est le nombre minimal de frontières à traverser pour aller de i à j. Quelques exemples de la contiguïté d'ordre 2 sont :

- *Double linear contiguity* : les éléments de la matrice de contiguïté vérifient $w_{i,j} = 1$ si l'observation j a au moins une frontière commune, à gauche ou à droite, avec la région contiguë de i et $w_{i,j} = 0$ sinon,
- *Double rock contiguity* : les éléments de la matrice de contiguïté vérifient $w_{i,j} = 1$ si l'observation j a au moins une frontière commune avec la région contiguë de i et $w_{i,j} = 0$ sinon,
- *k-nearest Neighbors* : les éléments de la matrice de contiguïté vérifient $w_{i,j} = 1$ si l'observation j a au moins une frontière commune avec la région contiguë d'ordre *(k-1)* de i et $w_{i,j} = 0$ sinon.

Condition de distance

Une autre condition possible pour déterminer l'ensemble des voisinages est la condition de distance. Les voisins d'une observation peuvent être définis comme

l'ensemble des observations localisées dans un rayon de *x* mètres autour de cette observation. Cette condition demande des informations précises sur la localisation telle que les coordonnées géographiques ou les coordonnées dans un plan cartésien afin de pouvoir calculer la distance entre les observations.

2.3.2. Matrice de poids

Une fois l'ensemble des voisinages défini, dans une deuxième étape, il faut déterminer les poids associés à chaque voisin. Ces poids peuvent être identiques pour le voisinage mais il est aussi possible d'accorder des poids important aux voisins proches et des poids faible aux voisins éloignés.

Matrice de poids standardisée (Row-standardized weight matrix)

La façon la plus simple de déterminer la matrice de poids spatiaux est d'attribuer un même poids à tous les voisins. Pour une observation donnée, supposons que *n* soit le nombre total de ses voisins. Le poids égal à l'inverse de nombre total de voisins, $\frac{1}{n}$, est attribué à tous les voisins de cette observation. La matrice *W* a la somme des éléments de chaque ligne égale à 1. Même si cette condition de poids standardisée est la plus simple à appliquer, elle présente un biais car dans la régression, cette condition accorde une importance plus grande aux observations ayant peu de voisins.

Matrice de poids binaire (Binary weight matrix)

Cette matrice accorde le même poids, égal à 1, à chaque voisin quelque soit le nombre total d'observations dans le voisinage. Les éléments de cette matrice sont donc soit 0, soit 1. La somme de chaque ligne donne le nombre d'observations dans le voisinage du point associé à cette ligne. Comparée à la matrice de poids standardisée, lors de la régression, cette condition accorde plus d'importance aux observations ayant beaucoup de voisins et moins d'importance aux observations ayant peu de voisins.

Matrice de poids binaire générale (General binary weight matrix)

Dans ce cas, le poids attribué à chaque observation voisine est variable et dépend de la distance séparant l'observation de son voisin. Les éléments de la matrice de poids générale sont souvent l'inverse de la distance séparant une observation de son voisin ou le carré de l'inverse de la distance séparant l'observation de son voisin. Le poids attribué est toujours une fonction décroissante de la distance; ainsi, deux observations proches présentent un degré de dépendance plus élevé par rapport deux observations plus éloignées. Cette condition de poids nécessite le calcul de la distance, et par conséquent, une information précise sur la localisation de l'observation est indispensable. La localisation peut être définie par les coordonnées cartésiennes ou les coordonnées géographiques. Selon l'objectif de l'étude, il existe plusieurs façons de calculer la distance entre deux observations : la distance à vol d'oiseau, la distance directe, la distance par la route ou la distance en temps de transport. Ces modes de calcul de la distance sont expliquées dans la section 2.2.2 du CHAPITRE I.

Les éléments de matrice W peuvent être plus élaborés selon l'objectif de recherche. Ci-dessous quelques exemples de matrices de distance souvent utilisées en étude immobilière.

- Fonction exponentielle négative : les éléments de la matrice de poids sont $w_{i,j} = e^{\alpha d_{ij}}$ avec d_{ij} la distance entre l'observation i et l'observation j, α est le paramètre à déterminer à priori,
- Fonction inverse de la distance : les éléments de la matrice de poids sont $w_{i,j} = \begin{cases} d_{ij}^{-\beta}, & d_{ij} < \bar{d} \\ 0, & d_{ij} \geq \bar{d} \end{cases}$ avec d_{ij} la distance entre l'observation i et j, β est le paramètre à déterminer à priori et $\bar{d}$ est aussi le seuil déterminé à priori indiquant la distance à partir de laquelle les observations i et j ne sont plus considérées comme voisins.

2.4. Méthodes d'estimation

En présence de dépendance spatiale, l'estimation par les moindres carrées ordinaires (MCO) donne des estimateurs biaisés et non efficients. Dans la première partie de cette section, les résultats de l'estimateur des MCO du modèle autorégressif spatial (SAR) et celui du modèle d'erreurs spatial (SEM) seront présentés. La méthode la plus élaborée par rapport aux MCO, habituellement proposée dans la littérature pour estimer la variable en présence de dépendance spatiale, est la méthode du maximum de vraisemblance. Cette méthode est présentée dans la deuxième partie. Une contrainte dans l'application de cette méthode reste cependant la complexité du calcul. Cette méthode est difficilement à mettre en œuvre si la taille de la base de données est importante. LeSage et Pace (2009) proposent donc la fonction de vraisemblance concentrée qui permet de réduire la dimension du programme d'optimisation. Cette fonction de log-vraisemblance concentrée est développée dans la troisième partie de cette section. Il existe des autres méthodes existantes mais elles ne sont pas présentées parce que cette section s'intéresse uniquement les méthodes utilisées en étude immobilière. Les méthodes parlées sont, par exemple, la méthode des moments généralisés (GMM) proposée par Kelejian et Prucha (1999) qui permet de simplifier le calcul de la maximisation de vraisemblance ou l'estimation par la méthode Bayésienne proposée par LeSage et Pace (2009) qui permet de faire le choix entre les différents modèles de la dépendance spatiale.

2.4.1. Estimation de moindre carrée ordinaire (MCO) et problème de dépendance spatiale

Gallo (2002) indique que lorsqu'une variable endogène présentant la dépendance spatiale est ignorée lors de la définition du modèle de régression mais cette dépendance présente dans le processus générateur des données, les estimateurs par les MCO seront biaisés et non convergents. Dans le cas où la dépendance spatiale des erreurs de l'estimation est ignorée lors de la spécification du modèle, même si les résultats de l'estimateur des MCO sont non biaisés mais ils ne sont pas efficients. Les estimateurs des MCO ainsi que leur propriété sont présentés dans cette section.

Modèle autorégressif spatiale pur (SAR)

Reprenons d'abord le modèle autorégressif spatiale (SAR) de la section 2.2.1 où la dépendance spatiale se présente dans les variables endogènes. Le modèle SAR se caractérise par le système d'équations suivant :

$$\begin{gathered} y = \rho W_{SAR} y + X\beta + \varepsilon \\ y = (I_n - \rho W_{SAR})^{-1}(X\beta + \varepsilon) \\ \varepsilon \sim N(0, \sigma^2 I_n) \end{gathered} \qquad \text{Eq. IV.31}$$

Si la dépendance des variables endogènes est ignorée lors de la spécification du modèle d'estimation ; la partie $W_{SAR}y$ est supprimée de l'Eq. IV.31. L'estimation par les moindres carrées ordinaires (MCO) donne l'estimateur, $\hat{\beta}$ qui est biaisé car :

$$\begin{gathered} \hat{\beta} = (X'X)^{-1}X'(I_n - \hat{\rho} W_{SAR})y \\ \hat{\beta} = (X'X)^{-1}X'y - \hat{\rho}(X'X)^{-1}X'W_{SAR}y \\ \hat{\beta} = \beta_0 - \hat{\rho}(X'X)^{-1}X'W_{SAR}y \end{gathered} \qquad \text{Eq. IV.32}$$

En plus, ces estimateurs ne sont pas convergents car la variable endogène décalée ($W_{SAR}y$) est corrélée avec l'erreur (ε) (Gallo (2002)).

$$\begin{aligned} E\{(W_{SAR}y)\varepsilon'\} &= E\{W_{SAR}(I_n - \rho W_{SAR})^{-1}X\beta\varepsilon' + W_{SAR}(I_n - \rho W_{SAR})^{-1}\varepsilon\varepsilon'\} \\ &= W_{SAR}(I_n - \rho W_{SAR})^{-1}E\{\varepsilon\varepsilon'\} \\ &= \sigma^2 W_{SAR}(I_n - \rho W_{SAR})^{-1} \\ &\neq 0 \end{aligned} \qquad \text{Eq. IV.33}$$

Modèle d'erreurs spatial (SEM)

Le modèle SEM présente l'autocorrélation spatiale des erreurs de l'estimation. En notant θ le coefficient de la dépendance spatiale des résidus et W_{SEM} la matrice de poids spatiaux, le modèle SEM se caractérise par :

$$\begin{gathered} y = X\beta + \varepsilon \\ \varepsilon = (I_n - \theta W_{SEM})^{-1}\epsilon \end{gathered} \qquad \text{Eq. IV.34}$$

L'Eq. IV.34 peut s'écrire comme :

$$y = X\beta + (I_n - \theta W_{SEM})^{-1}\epsilon$$
$$\epsilon \sim N(0, \sigma^2 I_n) \qquad \text{Eq. IV.35}$$

L'espérance et la variance des résidus de l'estimation (ε) s'écrivent:

$$E(\varepsilon) = 0$$
$$\Omega_\varepsilon = E(\varepsilon\varepsilon') = \sigma^2(I_n - \theta W_{SEM})^{-1}(I_n - \theta W'_{SEM})^{-1} \qquad \text{Eq. IV.36}$$

L'estimateur de β par les MCO du modèle SEM est non biaisé: $E(\hat{\beta}_{MCO} - \beta) = E((X'X)^{-1}X'\varepsilon) = (X'X)^{-1}X'E(\varepsilon) = 0$, mais il n'est pas efficient. En plus, l'estimation des MCO ne permet pas d'obtenir un estimateur consistent de θ. Le problème de non consistance de l'estimateur θ fait que les moindres carrées généralisés (MCG) ne permet pas d'obtenir l'estimation non biaisée de β.

Puisque ni les MCO ni les MCG ne permettent d'obtenir un estimateur non biaisé, convergent et efficient, nous devons faire recours aux méthodes d'estimation plus avancées. Gallo (2002) utilise la méthode du maximum de vraisemblance. LeSage et Pace (2009) proposent le maximum de vraisemblance concentré et la méthode bayésienne qui est basée sur le « *Markov Chain Monte Carlo* (MCMC) ». Kelejian et Prucha (1998) et Drukker, Prucha et Raciborski (2011) proposent la méthode de moindres carrés généralisés en deux étapes. Notre travail ne portera que sur le maximum de vraisemblance qui est la méthode la plus souvent utilisée dans en étude immobilière.

2.4.2. Maximisation de vraisemblance

En présence de dépendance spatiale, l'estimation par les moindres carrées ordinaires donne des résultats biaisés et non convergents. De plus, l'estimation par les moindres carrées généralisées n'est pas envisageable car un estimateur consistent du degré de dépendance spatial (ρ) est nécessaire. La méthode du maximum de vraisemblance est donc proposée pour estimer les paramètres du modèle de dépendance spatiale.

Reprenons le modèle spatial général (GSM) qui combine le modèle SAR et SEM. La dépendance spatiale est présente à la fois dans la partie des variables dépendantes

($W_{SAR}y$) et la partie des résidus ($W_{SEM}\varepsilon$). Ce modèle est défini par le système d'équations suivant :

$$y = X\beta + \rho W_{SAR}y + \varepsilon$$
$$\varepsilon = \theta W_{SEM}\varepsilon + \epsilon$$

d'où $$y = (I_n - \rho W_{SAR})^{-1}X\beta + (I_n - \rho W_{SAR})^{-1}(I_n - \theta W_{SEM})^{-1}\epsilon \quad \text{Eq. IV.37}$$
$$\epsilon \sim N(0, \sigma^2 I_n)$$

ρ et θ représentent les paramètres du degré de dépendance, W_{SAR} et W_{SEM} supposées connues sont les matrices de poids correspondant aux modèles SAR et SEM. Lorsque $\theta = 0$, l'Eq. IV.37 représente le modèle SAR et lorsque $\rho = 0$, l'Eq. IV.37 représente cette fois-ci le modèle SEM. Sous l'hypothèse de la normalité des erreurs $\epsilon \sim N(0, \sigma^2 I_n)$, les erreurs de l'estimation peuvent être décrites par :

$$\epsilon = (I_n - \theta W_{SEM})(y - \rho W_{SAR}y - X\beta) \quad \text{Eq. IV.38}$$

LeSage et Pace (2009) définissent le jacobien de cette transformation de la façon suivante :

$$J = det\left(\frac{\partial u}{\partial y}\right) = |I_n - \theta W_{SEM}||I_n - \rho W_{SAR}| \quad \text{Eq. IV.39}$$

La fonction log-vraisemblance est donc donnée par :

$$\ln L(y|\rho, \theta, \beta) = -\frac{n}{2}\ln(\pi\sigma^2) + \ln|I_n - \rho W_{SAR}| + \ln|I_n - \theta W_{SEM}| - \frac{\epsilon'\epsilon}{2\sigma^2} \quad \text{Eq. IV.40}$$

Les paramètres à estimer sont $\hat{\rho}$, $\hat{\theta}$, $\hat{\beta}$ et $\hat{\sigma}^2$. La démarche de la maximisation de la fonction de log-vraisemblance consiste à itérer les estimations de chaque paramètre conditionnellement à la valeur des autres paramètres.

$$\hat{\sigma}^2(\rho, \theta) = \frac{1}{n}y^{*\prime}(\rho, \theta)[I_n - X^*(\theta)[X^{*\prime}(\theta)X^*(\theta)]^{-1}X^{*\prime}(\theta)]y^*(\rho, \theta)$$
$$\hat{\beta}(\rho, \theta) = [X^{*\prime}(\theta)X^*(\theta)]^{-1}X^{*\prime}(\theta)y^*(\rho, \theta) \quad \text{Eq. IV.41}$$

avec

$$y^*(\rho, \theta) = (I_n - \theta W_{SEM})(I_n - \rho W_{SAR})y \quad \text{Eq. IV.42}$$

$$X^*(\theta) = (I_n - \theta W_{SEM})X$$

L'estimation par la maximisation de vraisemblance exige le calcul des déterminants de $|I_n - \theta W_{SEM}|$ et $|I_n - \rho W_{SAR}|$ à chaque itération.

Le calcul de la fonction de vraisemblance nécessite un temps et une capacité de calculs élevés, il faut créer la matrice de taille $n \times n$. En plus, il y a au moins quatre paramètres à estimer. LeSage et Pace (2009) indiquent que cette méthode nécessite un temps de calcul important, surtout lorsque la taille de l'échantillon est très élevée. Kelejian et Prucha (1999) indiquent que la méthode du maximum de vraisemblance n'est pas réalisable à cause de la complexité du calcul. Cette méthode est difficilement réalisable si la taille de la base de données est importante. En plus, le calcul de la fonction de vraisemblance devient encore plus compliqué si la matrice de poids n'est pas symétrique, ce qui est le cas si le voisinage est déterminé par la contiguïté d'ordre *k*. Plusieurs auteurs développent d'autres méthodes plus simple à mettre en œuvre comme la méthode du maximum de vraisemblance concentrée (LeSage et Pace (2009)) présentée ci-dessous.

2.4.3. Maximisation de vraisemblance concentrée

LeSage et Pace (2009) proposent la fonction log-vraisemblance *concentrée* qui permet de réduire la dimension du programme d'optimisation. Le problème d'optimisation multi variée de la fonction log-vraisemblance standard est remplacé par un problème d'optimisation uni-varié. La fonction de log-vraisemblance concentrée est développée différemment selon que la présence de la dépendance spatiale soit dans la partie des variables endogènes (modèle SAR) dans la partie des résidus (modèle SEM).

Modèle autorégressif spatiale (SAR)

En supposons que $\theta = 0$ dans l'Eq. IV.40, la fonction de log-vraisemblance standard du modèle SAR est :

$$\ln L(y|\rho,\beta) = -\frac{n}{2}\ln(\pi\sigma^2) + \ln|I_n - \rho W_{SAR}| - \frac{\epsilon'\epsilon}{2\sigma^2}$$
$$\epsilon = (y - \rho W_{SAR} y - X\beta) \qquad \text{Eq. IV.43}$$

La fonction log-vraisemblance concentrée permet de réduire la dimension de l'optimisation à un seul paramètre ρ, de sorte que les autres paramètres soient ensuite estimés en fonction de l'estimateur $\hat{\rho}$, définit par $\hat{\beta}(\hat{\rho})$ et $\hat{\sigma}^2(\hat{\rho})$. La fonction de log-vraisemblance concentrée du modèle SAR est définie par :

$$\ln L(\rho) = \kappa + \ln|I_n - \rho W_{SAR}| - \frac{n}{2}\ln\big(S(\rho)\big)$$
$$\text{avec } S(\rho) = e'_0 e_0 - 2\rho e'_0 e_d + \rho^2 e'_d e_d$$
$$e_0 = y - X\beta_0$$
$$e_d = Wy - X\beta_d \qquad \text{Eq. IV.44}$$
$$\beta_0 = (X'X)^{-1}X'y$$
$$\beta_d = (X'X)^{-1}X'Wy$$

κ est une constante qui ne dépend pas du paramètre ρ, $|I_n - \rho W_{SAR}|$ est le déterminant de la matrice $I_n - \rho W_{SAR}$. $e'_0 e_0$, $e'_0 e_d$ et $e'_d e_d$ sont des scalaires et les paramètres β_0 et β_d sont des vecteurs $k \times 1$. L'estimation du paramètre ρ par la méthode du maximum de vraisemblance concentré est plus rapide à mettre en œuvre car elle est conditionnelle aux valeurs $e'_0 e_0$, $e'_0 e_d$ et $e'_d e_d$. Une fois obtenu la valeur estimée de $\hat{\rho}$, nous pouvons estimer, par la suite, les coefficients du modèle ($\hat{\beta}$) et la matrice variance-covariance ($\widehat{\Omega}$) en fonction de $\hat{\rho}$ par les équations suivantes :

$$\hat{\beta} = \beta_0 - \hat{\rho}\beta_d$$
$$\hat{\sigma}^2 = n^{-1}S(\hat{\rho}) \qquad \text{Eq. IV.45}$$
$$\widehat{\Omega} = \hat{\sigma}^2[(I_n - \hat{\rho}W_{SAR})'(I_n - \hat{\rho}W_{SAR})]^{-1}$$

Modèle d'erreurs spatial (SEM)

En remplaçant $\rho = 0$ dans l'Eq. IV.40, la fonction de log-vraisemblance standard du modèle d'erreurs spatial (SEM) est la suivante :

$$\ln L(y|\theta,\beta) = -\frac{n}{2}\ln(\pi\sigma^2) + \ln|I_n - \theta W_{SEM}| - \frac{\epsilon'\epsilon}{2\sigma^2}$$
$$\epsilon = (I_n - \theta W_{SEM})(y - X\beta) \qquad \text{Eq. IV.46}$$

LeSage et Pace (2009) proposent la fonction log-vraisemblance concentrée de façon suivante :

$$\ln L(\theta) = \kappa + \ln|I_n - \theta W_{SAR}| - \frac{n}{2}\ln\big(S(\theta)\big)$$
$$\text{avec } S(\theta) = A_{yy}(\theta) - \beta(\theta)' A_{XX}(\theta)\beta(\theta)$$
$$A_{XX}(\theta) = X'X - \theta X'WX - \theta X'W'X + \theta^2 X'W'WX$$
$$A_{Xy}(\theta) = X'y - \theta X'Wy - \theta X'W'y + \theta^2 X'W'Wy \qquad \text{Eq. IV.47}$$
$$A_{yy}(\theta) = y'y - \theta y'Wy - \theta y'W'y + \theta^2 y'W'Wy$$
$$\beta(\theta) = A_{XX}(\theta)^{-1} A_{Xy}(\theta)$$

L'estimation du paramètre θ est plus simple que le maximum de vraisemblance standard parce que les éléments $A_{XX}(\theta)$, $A_{Xy}(\theta)$, $A_{yy}(\theta)$ sont pré-estimés. Une fois obtenu la valeur estimée de $\hat{\theta}$, les coefficients du modèle SEM ($\hat{\beta}$) et la matrice de variance-covariance ($\widehat{\Omega}$) peut être estimée de façon suivante :

$$\hat{\beta} = \beta_0 - \hat{\theta}\beta_d$$
$$\hat{\sigma}^2 = n^{-1} S(\hat{\theta}) \qquad \text{Eq. IV.48}$$
$$\widehat{\Omega} = \hat{\sigma}^2 \left[\left(I_n - \hat{\theta} W_{SAR}\right)'\left(I_n - \hat{\theta} W_{SAR}\right)\right]^{-1}$$

Même si le développement de la log-vraisemblance concentré permet de réduire la dimension du problème d'optimisation et augmente la rapidité de résolution du programme de maximisation, la taille importante de la matrice W de dimension $n \times n$ reste encore une difficulté à surmonter. LeSage et Pace (2009) proposent l'approche Bayésienne pour simplifier encore le calcul.

Le choix entre ces différentes méthodes d'estimation dépend de la complexité du modèle choisi et de la taille de la base de données. La méthode d'estimation présentée

devient plus compliquée si la dépendance spatiale se présente à la fois dans les variables endogènes et dans les résidus ou si le modèle inclut une autocorrélation entre les variables exogènes et les variables endogènes. Plusieurs logiciels statistiques propose un package qui permet d'analyser le modèle avec la dépendance spatiale et d'estimer le degré de corrélation spatiale. Drukker, Prucha et Raciborski (2011) proposent le package *spmat* et *gs2sls* attaché sur *Stata*. La fonction *spmat* est destinée à construire la matrice de poids et la fonction *gs2sls* permet d'estimer le modèle spatial avec la méthode de moindres carrés généralisés en deux étapes (A Generalized Spatial Two-Stage Least Squares – GS2SLS). Bivand (2012) propose le package *spdep* attaché à *R* qui permet de construire la matrice de poids (avec le fonction *knearneigh* et *dnearneigh*) et d'estimer les paramètres de corrélation spatiale avec plusieurs méthodes possibles. LeSage et Pace (2009) donnent les codes *Matlab* pour analyser les données spatiales. Comme chaque logiciel dispose d'une mémoire interne différente, le choix du logiciel dépendra non seulement de l'objectif de l'étude mais aussi de la taille de la base de données.

3. Économétrie spatiale et finance immobilière

La section 2 de ce chapitre décrit les différents éléments de la méthode de l'économétrie spatiale à savoir le processus spatial, les différents modèles de régression, la détermination de matrice de poids spatiaux et les méthodes d'estimation. Cette section a pour but de faire le lien entre la méthode d'économétrie spatiale et l'application en étude immobilière. La section 3.1 détaille les modèles et les méthodes d'estimation utilisés dans la littérature en étude immobilière. Les modèles d'économétrie spatiale utilisés en étude immobilière ne sont pas très variés ; on retrouve très souvent les trois modèles que sont le modèle SAR, le modèle SEM ou le modèle SDM. Puisque plusieurs modèles existent comme nous l'avons vu dans la section 2.2, la section 3.2 sera destinée à donner des exemples d'application des modèles d'économétrie spatiale en étude immobilière. Ce choix dépend normalement de la source de l'autocorrélation spatiale.

3.1. Étude économétrique spatiale immobilière

Dans la littérature en étude immobilière, l'économétrie spatiale semble plus connue et plus utilisée que la géostatistique. C'est grâce au développement des outils statistiques et économétriques qui permet à l'économétrie spatiale d'être mise en valeur. En plus, les progrès de l'informatique et le développement de logiciels économétriques permettent de faciliter l'application de cette méthode. Cette section présente une brève revue de la littérature sur l'approche issue de l'économétrie spatiale. Nous nous intéresserons aux modèles d'autocorrélation spatiale choisis, aux poids accordés aux observations et à la méthode d'estimation.

Le travail de Can (1990) est l'un des premiers articles en évaluation immobilière qui travaille sur l'économétrie spatiale. Il ajoute au modèle hédonique le terme qui permet de prendre en compte l'autocorrélation des prix immobiliers, ce qui donne le modèle SAR, et compare les résultats de ce modèle à ceux de la régression hédonique traditionnelle. Plusieurs matrices de poids spatiaux sont utilisées ; les poids sont l'inverse de la distance, l'inverse de la distance au carrée ou des poids binaires. La méthode d'estimation choisie est la méthode des MCO qui suppose que les erreurs de l'estimation ε sont indépendantes de la variable endogène décalée Wy. Cette hypothèse est très forte et peut conduire à des estimateurs biaisés, mais d'après Can (1990) la méthode du maximum de vraisemblance n'est pas réalisable à cause de la taille très large de sa base de données. Néanmoins, cette hypothèse d'indépendance est levée plus tard par Can (1992) qui utilise la méthode du maximum de vraisemblance pour estimer ces mêmes paramètres. Can et Megbolugbe (1997) proposent un indice de prix immobilier à partir des prix de biens résidentiels vendus à Miami. Ils choisissent le modèle SAR et deux matrices de poids spatiaux. La première matrice de poids utilise la condition de distance pour déterminer l'ensemble des voisinages et le poids égal à l'inverse de la distance est accordé à chaque couple d'observations. La deuxième matrice de poids emploie la contiguïté pour déterminer l'ensemble des voisinages et accorde un poids égal à l'inverse de la distance aux observations voisines. La méthode des moindres carrées ordinaires est choisie pour cette étude même si certaines études indiquent que l'estimation par les MCO donne des estimateurs biaisés et non convergent. Comme les études précédentes, Pace et Gilley (1997) utilisent aussi le modèle SAR mais leur matrice de poids diffère des autres.

Pour chaque observation, le poids accordé à une observation dans le voisinage correspond au rapport entre la distance et la distance maximale entre deux observations dans ce voisinage. La méthode utilisée pour estimer les paramètres est la méthode du maximum de vraisemblance. Brasington (1999) et Won Kim, Phipps et Anselin (2003) utilisent le modèle SAR mais leur méthode d'estimation diffèrent. Brasington (1999) utilise la méthode du maximum de vraisemblance. L'article plus récent de Won Kim, Phipps et Anselin (2003) utilisent deux méthodes d'estimation que sont la méthode du maximum de vraisemblance et le *spatial two stage least squares* S-2SLS proposé par Kelejian et Prucha (1998) et trouvent que les paramètres estimés par ces deux méthodes sont très similaires. Won Kim, Phipps et Anselin (2003) comparent les méthodes d'estimation et Wilhelmsson (2002) compare les modèles spatiaux et les matrices de poids. Wilhelmsson (2002) applique la méthode du maximum de vraisemblance pour estimer les paramètres des modèles SAR et SEM et compare les résultats. Il trouve que les résultats obtenus de ces deux modèles sont très proches. Il vérifie ensuite la variation des paramètres estimés selon la matrice de poids spatiaux. Six types de poids permettent de déterminer les matrices de poids spatiaux : (1) l'inverse de la distance, (2) l'inverse de la distance au carré, (3) l'inverse de la distance dans un voisinage d'une distance limite de 600 mètres (4) la contiguïté d'ordre 1, (5) la contiguïté d'ordre 4 et (6) la contiguïté en fonction des coordonnées géographiques[9]. Il trouve que les résultats sont du même ordre de grandeur, que le modèle (3) explique le mieux la variation des prix immobiliers et que le modèle (5) a le plus mauvais pourvoir explicatif. À partir du modèle spatial, certains auteurs ajoutent la dimension temporelle et développent le modèle spatio-temporel (STAR) (Clapp (2004); Pace, Barry, Clapp et Rodriquez (1998); Sun, Tu et Yu (2005); Tu, Yu et Sun (2004)). Tous les modèles spatio-temporels sont développés à partir du modèle SAR. Les travaux de Tu, Yu et Sun (2004) et Sun, Tu et Yu (2005) diffèrent des autres travaux par leur détermination de la matrice de poids spatiaux. Ils distinguent entre la matrice de poids spatiaux pour l'immeuble et la matrice de poids spatiaux pour le voisinage. Clapp (2004) développe la méthode de régression locale (Local Regression Model – LRM), une approche semi paramétrique pour estimer les paramètres du modèle STAR.

[9] Il crée le triangle reliant les coordonnées géographiques pour en déduire le voisinage.

Au vu de la revue de littérature présentée ci-dessus et malgré les différents modèles spatiaux existants, la majorité des études immobilières utilisent le modèle SAR ou le modèle SEM. Ce sont deux modèles de base qui prennent en compte la dépendance spatiale des prix immobiliers (la variable endogène) ou l'autocorrélation spatiale des erreurs de l'estimation hédoniques. Cependant, il est aussi possible que les caractéristiques physiques des biens voisins sont corrélées ou que l'autocorrélation se présente à la fois dans les prix immobiliers et leurs caractéristiques. Il existe encore plusieurs modèles non utilisés, mais qui peuvent l'être pour analyser l'autocorrélation des prix immobiliers. Le détail de ces différents modèles avec les applications possibles en étude immobilière se trouve dans la section 3.2. Concernant la matrice de poids spatiaux, le choix est plus varié. La contiguïté et la condition de distance sont, toutes les deux, utilisées pour déterminer l'ensemble des voisinages. Les poids accordés dans la majorité des cas est l'inverse de la distance. Ceci est peut-être dû au fait que l'inverse de la distance permet de mesurer l'accessibilité qui est une des caractéristiques principales en évaluation immobilière. L'explication plus détaillée sur le choix de la matrice de poids spatiaux est donnée dans la section 3.3. L'autre point important de l'économétrie spatiale est la méthode d'estimation. Dans la majorité des cas, la méthode choisie est le maximum de vraisemblance, malgré sa complexité de calcul. Les méthodes plus sophistiquées qui permettent de réduire le temps de calcul comme la méthode du maximum de vraisemblance concentré, la méthode des moments généralisés (GMM) ou l'approche Bayésienne proposé par LeSage et Pace (2009) ne sont pas encore utilisées en étude immobilière.

3.2. Choix du modèle économétrique spatial

La prise en compte de l'autocorrélation spatiale dans les modèles économétriques peut s'effectuer de plusieurs manières : par des variables spatiales décalées, endogènes ou exogènes, ou par une autocorrélation spatiale des erreurs (Gallo (2002)). Ainsi, les modèles spatiaux présentés dans la section 2.2 diffèrent l'un de l'autre selon les variables présentant la structure spatiale. Cette différence est due aux différentes sources de dépendance spatiale prises en compte dans chaque modèle. Cette section est destinée à

décrire les modèles spatiaux selon leur source de dépendance spatiale et à donner des exemples d'application dans un contexte immobilier.

Le modèle autorégressif spatial (SAR)

Le modèle spatial le plus basique est le modèle autorégressif spatial (SAR) où l'autocorrélation spatiale de la variable endogène est prise en compte. Cressie (1991) indique que le modèle SAR est utilisé dans le cas où l'effet de diffusion (*spillover effect*) cause la dépendance spatiale de la variable endogène ; un changement a une influence directe sur une observation et crée une influence indirecte sur les observations voisines. Le modèle SAR est le modèle le plus utilisé en étude immobilière car le processus d'évaluation de bien immobilier crée la dépendance des prix. Le vendeur se renseigne notamment auprès de ses voisins pour déterminer le prix de vente de son bien. Quant à l'acheteur, il s'intéresse à la valeur des biens localisés autour du bien qu'il veut acheter afin d'estimer le prix d'achat le plus raisonnable. La valeur d'un bien est donc calculée en prenant en compte la valeur estimée des biens voisins comme référence et y ajoutant un surplus correspondant aux caractéristiques spécifiques du bien ainsi qu'une plus-value liée à la préférence de l'acheteur. Même chez les professionnels, l'expert immobilier estime la valeur d'un bien en fonction des valeurs des biens voisins.

Remarquons que dans le modèle SAR, l'information se transfère en passant uniquement par le prix. Une hausse du prix d'un bien a une influence sur les prix de biens voisins. Par exemple, la rénovation de la façade d'un immeuble a une influence directe positive sur le prix de vente des appartements localisés dans cet immeuble. Et cette augmentation de prix a une influence sur le prix des appartements localisés dans les immeubles voisins. La rénovation de la façade d'un bien a donc un effet positif indirect sur les prix des biens voisins en passant par le processus d'évaluation du bien. Un autre exemple est l'existence d'un centre commercial qui augmente le prix des biens résidentiels les plus proches de ce centre. Grâce à l'effet de diffusion, cette hausse de prix a une influence sur le prix des appartements situés plus loin. Ainsi, l'impact de l'existence du centre commercial sur le prix immobilier se diffuse mais l'importance de cet impact dépend négativement de la distance entre le centre commercial et l'appartement.

Le modèle SAR peut aussi être utilisé dans une étude du développement urbain, pour étudier l'impact sur le prix immobilier lié à la modernisation du quartier ou la création du service de transport public. Cressie (1991) illustre l'exemple d'une étude des variations de prix immobiliers liées à la création d'une nouvelle autoroute reliant une commune au centre-ville. Cette nouvelle autoroute améliore la facilité de déplacement non seulement pour la commune qu'elle dessert directement, mais aussi ses communes voisines. Même si certaines communes ne sont pas directement desservies par cette autoroute, cette infrastructure permet de réduire la durée totale du déplacement entre ces communes et le centre-ville. Ainsi, la nouvelle autoroute a une influence directe sur les prix immobiliers de la commune qu'elle dessert et une influence indirecte sur les prix immobiliers des communes voisines. La création de l'autoroute diffuse une influence positive sur les prix immobiliers des communes où cette autoroute passe et crée une dépendance spatiale entre les prix immobiliers parmi ces différentes communes.

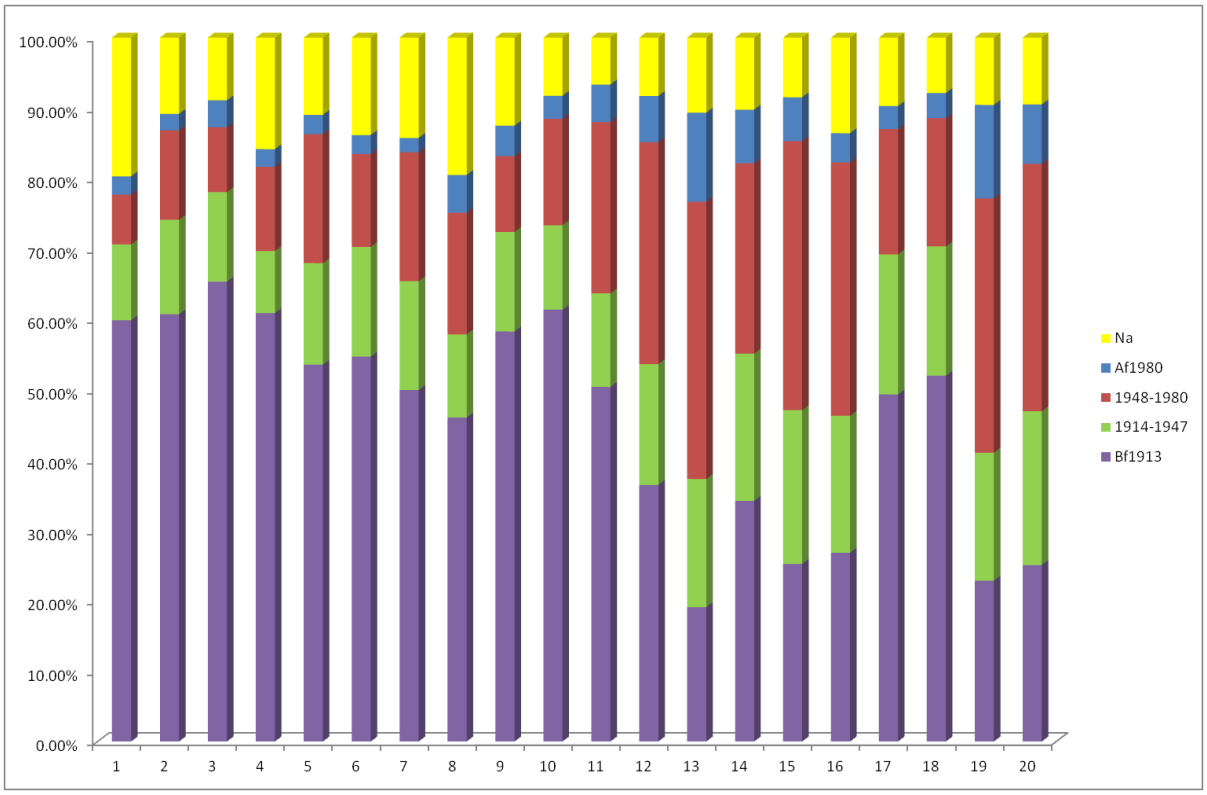
Le modèle lag-X spatial (SLX)

Le modèle SLX prend en compte la dépendance spatiale des variables exogènes. Deux raisons peuvent être expliquées la dépendance spatiale des variables explicatives.

Premièrement, la ressemblance des biens voisins. Les biens voisins ont souvent les mêmes caractéristiques : ils sont construits dans la même période, ils ont la même structure et le même âge. Cette ressemblance crée l'autocorrélation spatiale des caractéristiques physiques des biens ; ainsi les variables explicatives sont corrélées. Prenons l'exemple de l'âge d'un bien qui présente un important pouvoir explicatif dans l'évaluation des biens immobiliers. Les immeubles localisés dans un même quartier sont souvent construits dans la même période. Il est donc possible que la période de construction présente la dépendance spatiale. La Figure IV.1 présente le pourcentage des périodes de constructions des appartements vendu en 2007 dans les 20 arrondissements de Paris. Le pourcentage des appartements construits avant 1913 est important pour les 11 premiers arrondissements de Paris. Les appartements construits entre 1948-1980 se trouvent en majorité dans les 12ème au 16 ^{ème} arrondissements. Ainsi donc, si une étude spatiale a pour l'objectif de mesurer le pouvoir explicatif de l'âge du bien sur le prix immobilier, le modèle intégrant la dépendance spatiale des variables explicatives doit être

choisi. Un autre exemple est la qualité de l'immeuble. C'est une variable explicative corrélée à tous les appartements de l'immeuble. La qualité de la maison peut aussi présenter la dépendance spatiale. Afin de déterminer la valeur d'un bien, la qualité de la maison peut être considérée comme variable explicative, mais le modèle peut aussi prendre en compte la qualité des maisons voisines. En effet, une maison qui est entourée par des maisons bien entretenues a un prix plus élevé qu'une maison de mêmes caractéristiques mais qui est entourée par des maisons en mauvaise qualité. La qualité de la maison est donc une des variables explicatives présentant la dépendance spatiale.

Figure IV.1 Le pourcentage des périodes de constructions des appartements vendu en 2007 dans les 20 arrondissements de Paris



La deuxième explication de la dépendance spatiale des variables explicatives est l'influence d'une externalité. Une externalité a normalement une influence sur une zone ou un quartier, et donc les prix des biens immobiliers localisés dans cette zone sont tous influencés par cette externalité. Par exemple, la valeur d'un bien immobilier dépend du

taux de criminalité de l'arrondissement où se localise le bien mais elle peut dépendre aussi du taux de criminalité des autres arrondissements voisins. La nuisance sonore due à un grand boulevard a un effet négatif sur le prix des appartements situés le long de ce boulevard. Si le taux de criminalité ou la nuisance sonore sont considérés comme des variables explicatives, alors elles présenteront la dépendance spatiale.

Remarquons que le modèle SLX est différent du modèle SAR. Le modèle SAR considère que la dépendance spatiale se présente dans la partie des variables endogènes et donc dans les prix de biens immobiliers. Ce modèle considère que la variation dans une variable explicative a une influence directe sur la valeur du bien et c'est la variation de prix de ce bien qui a un effet indirect sur le prix des biens voisins. L'influence se transmet uniquement par le prix du bien. Par contre, le modèle SLX considère que la caractéristique d'un bien a une influence directe sur le prix du bien voisin.

Le modèle d'erreurs spatial (SEM)

Le modèle SEM est un autre modèle souvent utilisé en étude immobilière. Ce modèle prend en compte la dépendance spatiale des erreurs de l'estimation. Deux motivations souvent mentionnées pour expliquer l'autocorrélation des erreurs sont la variable omise lors de la définition de modèle et l'hétérogénéité spatiale. Premièrement, lors de la définition du modèle régression, certaines variables ne sont pas prises en compte comme variable explicative. Cela peut être dû à un manque d'information ou à une mauvaise spécification du modèle. Si cette variable explicative omise ne présente pas de dépendance spatiale, ce modèle donnera des estimateurs non biaisés. Mais si cette variable présente la dépendance spatiale, les erreurs de l'estimation seront corrélées. Reprenons l'exemple du taux de criminalité du modèle SLX. Le taux de criminalité de la région est une variable explicative nécessaire pour évaluer un bien immobilier et il présente le caractère de dépendance spatiale. Si cette variable n'est pas inclue dans le modèle d'évaluation ou si cette information n'est pas disponible dans la base de données, la valeur correspondant à cette variable est contenue dans la partie des erreurs de l'estimation. Ces erreurs sont donc spatialement corrélées.

LeSage et Pace (2009) soulignent que le modèle SEM peut aussi être développé pour résoudre le problème d'hétérogénéité. L'un des caractéristiques physiques d'un logement prises en compte dans l'évaluation de son prix c'est par exemple l'existence de terrasse. La variable explicative « existence de terrasse » a une influence sur la valeur de logement qui diffère selon la localisation du bien. Cette variable est beaucoup plus valorisée pour un logement en ville que pour un logement rural. Si l'on suppose que cette caractéristique a la même influence sur les logements en ville et les logements ruraux, cela crée un problème d'hétérogénéité qui conduit à une autocorrélation spatiale des erreurs de l'estimation.

Le modèle SEM et le modèle SLX permettent tous les deux de prendre en compte la dépendance spatiale causée par l'influence d'une externalité, mais la différence est que tous deux ne prennent pas en compte cette externalité lors de la spécification du modèle régression. Pour le modèle SLX, l'externalité est mesurable et la variable existe dans la base de données. Cette variable explicative est donc ajoutée au modèle de régression comme une des variables explicatives. Par contre, le modèle SEM considère que l'effet de l'externalité est non mesurable ou cette variable n'existe pas dans la base de données. Il est donc impossible d'ajouter cette variable explicative dans le modèle d'estimation. L'impact de l'externalité se présente donc dans les erreurs de l'estimation.

Le modèle de Durbin spatial (SDM)

Le modèle de Durbin spatial (SDM) prend à la fois en compte la dépendance spatiale de la variable endogène causée par le processus d'évaluation des biens immobiliers et celle des variables exogènes causée par la ressemblance de biens voisins. Le modèle SDM peut être considéré comme une extension du modèle SAR qui combine le modèle SAR et le modèle SLX. Ce modèle permet de prendre en compte les deux effets direct et indirect de la dépendance spatiale. Le premier effet est l'effet *indirect* sur le prix : une variable explicative a une influence sur le prix qui, par le processus d'évaluation, diffuse cet effet sur le prix de bien voisin. Le deuxième est l'effet *direct* sur le prix : comme les caractéristiques des biens voisins sont aussi prises compte dans la régression, les caractéristiques d'un bien ont un effet direct sur le prix d'un bien voisin.

Reprenons l'exemple précédent de la rénovation de la façade d'un immeuble. Selon le modèle SAR, cette rénovation a une influence directe sur le prix des appartements localisés dans cet immeuble. Ensuite, cette variation du prix de ces appartements va se transmettre sur les prix des appartements localisés dans un immeuble voisin en passant par le processus d'évaluation. Dans le cas du modèle SDM, nous considérons que cette rénovation améliore la vue depuis la rue et donc la caractéristique de localisation de tous les biens de cette rue. Par conséquent, en plus de l'effet indirect sur le prix, cette rénovation a aussi une influence directe sur le prix de biens voisins. Un autre exemple de la qualité de l'appartement, donné pour expliquer le modèle SLX. Pour le modèle SDM, nous considérons que pour déterminer le prix d'un appartement, l'acheteur prend en compte, à la fois, le prix moyen des appartements voisins comme le prix de référence et la qualité de bien voisin.

Les modèles SDM et SAR permettent d'inclure les influences directe et indirecte liées d'une variable explicative. Le modèle SDM distingue les influences directe et indirecte du changement de variable explicative dans le modèle de valorisation immobilière. L'importance accordée à ces deux influences peut être différente selon les deux coefficients de corrélation spatiale. Le SAR ne prend en compte que l'influence indirecte ; le changement d'une variable explicative a une influence indirecte sur le prix de bien voisin en passant uniquement par le processus d'évaluation.

Le modèle spatial général (GSM)

Le modèle de dépendance spatiale le plus généralisée est le modèle GSM. Ce modèle intègre à la fois la dépendance spatiale dans la partie de variable endogène (le prix immobilier) et la partie de l'erreur de l'estimation (l'effet non mesurable). Ce modèle peut être utilisé pour tester la présence de corrélation spatiale dans la variable endogène et, en même temps, de pouvoir vérifier si certaines variables exogènes présentent le caractère de dépendance spatiale. Si le coefficient de corrélation des résidus est significativement différent de zéro, cela signifie bien la présence de la dépendance spatiale dans certaines variables explicatives. Le modèle est alors mal spécifié. Une fois ajoutée la variable exogène présentant la dépendance spatiale, cette autocorrélation spatiale des erreurs devrait disparaitre.

Le modèle autorégressif moving average spatial (SARMA)

Le modèle SARMA ressemble au modèle GSM, la dépendance spatiale présente dans la partie de variable endogène et la dépendance des erreurs de l'estimation. Entre ces deux modèles, une différence est que le GSM considère que les erreurs de l'estimation suivent un processus autorégressif, alors que le SARMA considère qu'elles suivent le processus moyen mobile. Le processus autorégressif est un processus de dépendance à mémoire « infinie ». Le modèle GSM est utilisé pour étudier l'effet de diffusion global. Inversement, le processus moyen mobile est un processus de dépendance à mémoire « finie ». Le modèle SARMA permet d'étudier l'effet de diffusion local (Fingleton (2008)).

En étude immobilière, si les variables omises sont, par exemple, la qualité de l'air, la qualité du quartier ou l'existence de l'espace vert, les erreurs sont un processus autorégressif. En effet, ces variables manquantes ont une influence sur l'ensemble des observations du quartier. Par contre, si les variables explicatives omises sont, par exemple, la qualité de l'immeuble, la sécurité de l'entrée de l'immeuble, l'influence sera uniquement sur les prix des appartements situés dans cet immeuble. Un autre exemple : si la variable omise est l'indicateur du risque inondation alors l'impact négatif sera uniquement sur le prix uniquement pour les biens situés dans la zone risquée. Dans ces cas, les erreurs suivent un processus moyen mobile.

Le modèle de Durbin et d'erreurs spatial (SDEM)

Le modèle SDEM est la combinaison du modèle de Durbin et d'erreurs spatial (SDM) et du modèle d'erreurs spatial (SEM). Le modèle prend en compte, à la fois, la dépendance spatiale de la variable exogène et celle des erreurs d'estimation. Ce modèle considère que la dépendance spatiale provient uniquement des caractéristiques des biens et que ces caractéristiques ne sont pas toutes observables, certaines étant prises en compte dans la définition du modèle de régression et d'autres non. Ces variables spatiales omises lors de la spécification de modèle donnent la dépendance spatiale des erreurs de l'estimation. Ce modèle apparaît comme le moins réaliste en étude immobilière car il ne prend pas en compte la dépendance spatiale des prix immobiliers qui est un aspect essentiel des études immobilières. Cependant, ce modèle peut être utilisé pour analyser la

dépendance spatiale dans le cas où la distribution spatiale des données est très étalée. Par exemple dans une étude des variations de prix de biens localisés dans une commune ayant peu de biens ou des biens dispersés dans la commune, il est peu raisonnable de considérer que les prix de tels biens sont corrélés à cause du processus d'évaluation. Mais la décision d'augmentation de taxes dans une commune a une influence sur le prix de tous les biens localisés dans cette commune. Par ailleurs, il est possible qu'une augmentation de taxes soit destinée à améliorer la qualité de la commune. Ainsi, si cette variable n'est pas prise en compte lors de la spécification du modèle de régression, les erreurs de l'estimation seront corrélées.

Pour simplifier la comparaison et le choix entre les différentes modèles, le Tableau IV.1 présente le résumé de tous les modèles. Ce tableau indique l'équation de régression de chaque modèle, la partie présentant la dépendante spatiale, la source de la dépendance et l'utilisation du modèle dans l'étude immobilière.

Tableau IV.1 : Différents modèles d'économétrie spatiale classés selon la façon de prendre en compte l'autocorrélation spatiale, la source de l'autocorrélation spatiale et l'utilisation en étude immobilière

Modèles	Equations	Parties dépendantes	Utilisation en étude immobilière et source de la dépendance
Modèle autorégressif spatial (SAR)	$y = \alpha\iota_n + \rho W y + X\beta + \varepsilon$ $\boldsymbol{y = (I_n - \rho W)^{-1}(\alpha\iota_n + X\beta) + (I_n - \rho W)^{-1}\varepsilon}$ $\varepsilon \sim N(0, \sigma^2 I_n)$	– Variable endogène	– Valeur d'un bien immobilier corrélée aux valeurs de biens voisins. – Effet de diffusion (*Spillover effect)* – Processus d'évaluation de biens
Modèle de variables exogènes décalées (SLX)	$\boldsymbol{y = \alpha\iota_n + X\beta + WX\gamma + \varepsilon}$ $\varepsilon \sim N(0, \sigma^2 I_n)$	– Variable exogène	– Caractéristiques de biens corrélées – Ressemblance des biens voisins – Effet d'une externalité observable
Modèle d'erreurs spatial (SEM)	$y = \alpha\iota_n + X\beta + \varepsilon$ $\varepsilon = \theta W\varepsilon + \epsilon$ $\boldsymbol{y = \alpha\iota_n + X\beta + (I_n - \theta W)^{-1}\epsilon}$ $\epsilon \sim N(0, \sigma^2 I_n)$	– Erreur de l'estimation (processus autorégressif)	– Variable omise lors de la spécification du modèle d'estimation – Effet d'une externalité non observable – Hétérogénéité spatiale
Modèle de Durbin spatial (SDM)	$\boldsymbol{y = \rho W y + X\beta - \rho WX\beta + \varepsilon}$ $(I_n - \rho W)y = (I_n - \rho W)X\beta + \varepsilon$ $\varepsilon \sim N(0, \sigma^2 I_n)$	– Variable endogène – Variable exogène	– Processus d'évaluation et Ressemblance de biens voisins – Influence directe et indirecte du changement des caractéristiques des biens voisins.

Modèles	Equations	Parties dépendantes	Applications en immobilier
Modèle spatial général (GSM)	$y = \alpha\iota_n + X\beta + \rho W_{SAR} y + \varepsilon$ $\varepsilon = (I_n - \theta W_{SEM})^{-1}\epsilon$ $\boldsymbol{y = (I_n - \rho W_{SAR})^{-1}(\alpha\iota_n + X\beta) + (I_n - \rho W_{SAR})^{-1}(I_n - \theta W_{SEM})^{-1}\epsilon}$ $\epsilon \sim N(0, \sigma^2 I_n)$	– Variable endogène – Erreur de l'estimation (processus autorégressive)	– Effet de diffusion et effet global d'une externalité non observable – Verifier la spécification du modèle
Modèle autorégressif et moyenne mobile spatial (SARMA)	$y = \alpha\iota_n + X\beta + \rho W_{SAR} y + \varepsilon$ $\varepsilon = (I_n - \theta W_{SEM})\epsilon$ $\boldsymbol{y = (I_n - \rho W_{SAR})^{-1}(\alpha\iota_n + X\beta) + (I_n - \rho W_{SAR})^{-1}(I_n - \theta W_{SEM})\epsilon}$ $\epsilon \sim N(0, \sigma^2 I_n)$	– Variable endogène – Erreur de l'estimation (processus moyenne mobile)	– Effet de diffusion et effet local d'une externalité non observable
Modèle de Durbin et d'erreurs spatial (SDEM)	$y = \alpha\iota_n + X\beta + WX\gamma + \varepsilon$ $\varepsilon = (I_n - \theta W)^{-1}\epsilon$ $\boldsymbol{y = \alpha\iota_n + X\beta + WX\gamma + (I_n - \theta W)^{-1}\epsilon}$ $\epsilon \sim N(0, \sigma^2 I_n)$	– Variable exogène – Erreur de l'estimation (processus autorégressive)	– Ressemblance de biens voisins dans le cas de distribution spatiale très étalée

y : la matrice $n \times 1$ des variables endogènes, X : la matrice des variables explicatives de dimension $n \times k$, ι_n : le vecteur identité de dimension $n \times 1$ ne contient que la valeur 1, I_n : la matrice identité $n \times n$, ε et ϵ : le vecteur dimension $n \times 1$ des résidus de l'estimation, α et β : les coefficients de régression, ρ : le degré de la dépendance spatiale de variable endogène, γ : le degré de la dépendance spatiale de variable exogène, θ: le degré de la dépendance spatiale de l'erreur de l'estimation, W : la matrice de poids spatiaux de dimension $n \times n$ avec $w_{ij} > 0$ quand l'observation *j* est dans le voisinage de l'observation *i* et $w_{ij} = 0$ sinon.

3.3. Choix de la matrice de poids spatiaux

Dans la première étape de la détermination de la matrice de poids spatiaux, il faut déterminer l'ensemble des voisinages. Deux conditions habituellement utilisée sont la contiguïté et la condition de distance. Le choix entre ces deux conditions dépend du type des données, du nombre d'observations et de la distribution des observations dans l'espace

La matrice de contiguïté est la plus utilisée grâce à sa simplicité, mais le choix du degré de voisinage reste encore ambigu. Si la base de données est très dense, la contiguïté d'ordre 1 n'est pas suffisante. La valeur d'un bien immobilier est corrélée non seulement à la valeur des biens voisins localisés sur le coté mais aussi à la valeur des biens qui se trouve en face. Si la base de données est très étalée, la contiguïté d'ordre 1 pourrait considérer, comme voisins, un couple d'observations qui sont trop éloignés. Par exemple, s'il existe une seule observation localisée sur une île isolée, il n'est pas raisonnable de considérer que cette observation est corrélée à celle localisée sur la terre ferme. Cette condition pose donc problème s'il existe des observations isolées.

Comme l'autocorrélation spatiale est souvent définie en fonction de la distance séparant les observations, la condition de distance respecte parfaitement cette définition et permet de mieux cibler le voisinage dans le cas où la base de données est dense et régulière. Il y'a un problème si la base de données est irrégulière et étalée parce que le nombre d'observations dans le voisinage peut être différent d'une observation à l'autre. De plus, il existe le risque d'avoir une observation isolée, sans voisins, si la distance seuil n'est pas suffisamment élevée. Le choix de cette distance limite est un choix délicat qui est en fonction de la distribution des données et du sujet de recherche. Si la distance seuil est faible, cette condition permet de ne prendre en compte que le voisinage d'ordre 1 ou d'ordre 2 mais il peut exister des observations isolées sans voisin. Inversement, si la distance limite choisie est élevée, cette condition permet d'éliminer le problème d'observations isolées, mais le nombre de voisins peut être différent d'une observation à une autre et même être trop important pour certaines observations.

Après avoir déterminé l'ensemble des voisinages, il faut dans une deuxième étape déterminer le poids accordé à chaque voisin. Gallo (2002) mentionne que le choix de

matrice de poids spatiaux joue un rôle important dans le test de la dépendance spatiale. Le rejet ou l'acceptation de l'hypothèse nulle (l'hypothèse nulle est normalement l'absence de la dépendance) peut différer selon la définition de l'ensemble des voisinages. Le choix de cet ensemble dépend principalement de l'objectif de l'étude. Si la variable étudiée est un flux ou un mouvement, la matrice de distance est plus appropriée. Par contre, si cette variable correspond à un indicateur ou un niveau, la matrice qui semble la plus convenable est la matrice de contiguïté. Il est donc nécessaire de vérifier la robustesse du test par rapport à la définition de voisinage choisie.

4. Conclusion

Ce chapitre a permis de décrire les différentes étapes de la méthode de l'économétrie spatiale, qui est la méthode la plus utilisée en étude immobilière. Trois points constituent les étapes les plus importantes de la méthode de l'économétrie spatiale. Le premier point est le choix du modèle spatial. Malgré l'existence d'une variété de modèles spatiaux dans la littérature sur la statistique spatiale, seuls deux modèles spatiaux sont utilisés pour étudier la dépendance spatiale des prix immobiliers. Par conséquent, un des objectifs principaux de ce chapitre a été de décrire les modèles spatiaux selon la partie présentant l'autocorrélation spatiale à savoir la variable endogène, la variable exogène ou les erreurs de l'estimation, d'expliquer la source de l'autocorrélation de chaque variable et de donner des exemples d'application des modèles en étude immobilière. Le deuxième point important est le choix de la matrice de poids. Il faut d'abord choisir la condition, telle que la contiguïté ou la condition de distance, qui permet de déterminer l'ensemble des voisinages. Ensuite, le poids accordé aux observations dans le voisinage peut être le poids binaire, le poids standardisé ou l'inverse de la distance, selon l'objectif de l'étude, le type de données et la distribution spatiale des données. Le troisième point est la méthode d'estimation. Dans le cas de l'autocorrélation spatiale de la variable endogène, l'estimateur des MCO est biaisé et non convergent. Il faut alors recours à de méthodes d'estimation plus sophistiquées. La méthode la plus utilisée est celle du maximum de vraisemblance. Mais si cette méthode permet d'obtenir un estimateur non biaisé et efficient, elle nécessite un temps de calcul très long lorsque la base de données est de grande taille, et un logiciel de mémoire importante est indispensable de sorte à pouvoir

garder en mémoire tous les éléments de la matrice. En même temps, plusieurs méthodes d'estimation sont développées afin de simplifier le calcul. On distingue ainsi la méthode du maximum de vraisemblance, la méthode des moments généralisés, l'approche bayésienne ou le *Generalized Spatial Two-Stage Least Squares.*

CHAPITRE V DEGRE DE CORRELATION ET QUARTIER DOMINANT DU MARCHE IMMOBILIER FRANÇAIS

1. Introduction

Quel est le quartier résidentiel dominant de la ville ? Comment peut-on déterminer le quartier dominant ?

Si la ville en question correspond à une ville mono-centrique, la réponse est sans doute le centre ville. En revanche, si cette ville correspond à une ville polycentrique, la réponse dépend des connaissances géographiques ou historiques. Comment définir alors économétriquement le quartier dominant ? Il est difficile de définir un quartier dominant à partir d'un seul indicateur administratif existant. Plusieurs critères peuvent être pris en compte pour déterminer le quartier dominant : la qualité de l'arrondissement, le niveau de richesse des habitants, le nombre de commerces, l'intensité de la desserte en transports en commun, la sécurité, le niveau des prix immobiliers et même la présence de service public (la crèche, le gymnase, la piscine…).

Afin de répondre à ces questions, nous établissons un lien entre l'effet de diffusion (*spillover effect*) et la corrélation spatiale des prix immobiliers. La majorité des études immobilières indique que le processus d'évaluation d'un prix immobilier par son propriétaire entraine le problème de dépendance spatiale des prix. Pour déterminer la valeur de son bien, le propriétaire peut se renseigner, soit auprès de l'expert immobilier qui donne une estimation du prix basée sur la valeur des transactions sur les biens voisins, soit directement auprès des propriétaires des biens voisins. La moyenne pondérée des prix de vente des biens voisins est utilisée comme prix de référence pour déterminer la valeur de son bien. Il existe donc un processus d'interaction entre les valeurs des biens. Lorsque les prix immobiliers sont spatialement corrélés, une hausse de la valeur d'un bien se diffuse sur les biens voisins en passant par le processus d'évaluation.

Considérons une unité d'étude plus grande, soit le quartier ou l'arrondissement. D'après le principe de diffusion, le développement d'un quartier, par exemple en tant que quartier des affaires, conduit donc à une hausse du prix moyen des biens localisés dans ce quartier. En passant par le processus d'évaluation, cette hausse de prix se diffuse sur les prix moyens des autres quartiers alentours.

Nous utiliserons le degré de corrélation spatiale comme mesure de l'influence d'un quartier sur d'autres quartiers. Le quartier dominant est défini comme le quartier qui joue un rôle directeur sur les prix immobiliers de la ville, l'évolution des prix de ce quartier se diffusant sur les prix immobiliers des autres quartiers de la ville. Le quartier dominant est déterminé de la façon suivante. Lorsque les observations d'un quartier sont enlevées de la base de données, si le niveau de corrélation spatiale estimé à partir des observations restantes baisse significativement comparé au niveau de corrélation spatiale estimé à partir de l'ensemble des données, ce quartier sera considéré comme un quartier dominant. Cette réduction du niveau de corrélation spatiale signifie que ce quartier présente un rôle directif dans l'effet de diffusion et les prix immobiliers des autres quartiers sont partiellement corrélés avec les prix immobiliers de ce quartier dominant. Autrement dit, si la corrélation spatiale des prix immobiliers augmente lorsque les données d'un quartier sont prises en compte dans le modèle d'estimation, cela signifie que ce quartier présente un rôle directeur dans la détermination de la valeur des biens immobiliers de la ville.

Inversement, il est possible que le niveau de corrélation spatiale augmente significativement lorsque les observations d'un quartier sont enlevées de la base de données. Autrement dit, lorsque les données d'un quartier sont prises en compte dans le modèle d'estimation, si la corrélation spatiale des prix immobiliers baisse, alors ce quartier présente une caractéristique spécifique qui ne se trouve pas dans les autres quartiers. C'est à cause de cette caractéristique que le prix des biens immobiliers de ce quartier est peu corrélé aux prix des autres quartiers. Ce quartier peut être défini comme un quartier spécifique de la ville.

La détermination d'un quartier dominant est utile pour les acheteurs, les vendeurs et les investisseurs en immobilier. Dans le cas d'une grande ville ou d'une ville polycentrique, les experts immobiliers qui connaissent bien la ville peuvent indiquer le quartier dominant, mais le consommateur ou l'investisseur qui ne connaît pas bien le terrain a besoin d'un critère pour le déterminer. De plus, dans le cas de l'estimation de la valeur de biens immobiliers, le prix moyen des biens immobiliers localisés dans ce quartier dominant peut être utilisé comme un prix de référence. L'indice de prix

immobilier de ce quartier dominant permet de donner une information sur l'évolution du marché de la ville.

Nous utilisons le modèle autorégressif spatial (SAR) pour analyser les interactions entre les prix des appartements de différentes villes en France. La base de données comporte les douze agglomérations urbaines [10] de France (Bordeaux, Lille, Lyon, Marseille, Montpellier, Nantes, Nice, Orléans, Paris, Rennes, Strasbourg, Toulouse), et les trois communes de la petite couronne de Ile-de- France : Hauts-de-Seine, Seine-Saint-Denis et Val-de-Marne. Le nombre total d'observations traitées est environ 210 000 transactions immobilières en 1998 et 2007. Cette étude se sépare en deux niveaux selon l'unité de l'étude : l'étude au niveau agrégé et l'étude au niveau de la transaction.

Les objectifs de ce travail sont premièrement de mettre en exergue la présence de l'autocorrélation spatiale des prix immobiliers et mesurer le degré de dépendance spatiale dans chaque ville. Le deuxième objectif, qui est aussi le principal objectif de ce travail, est de déterminer le quartier dominant du marché immobilier dans chaque ville. Ce rôle dominant est examiné en mesurant, pour chaque quartier, le degré de l'autocorrélation spatiale avec et sans les données de ce quartier. Une baisse significative de la corrélation spatiale après exclusion des données d'un quartier donné indique que ce quartier peut être défini comme un quartier dominant du marché immobilier de la ville et, inversement, une hausse significative de la corrélation spatiale après exclusion des données d'un quartier donné indique que ce quartier présente une caractéristique spécifique par rapport aux autres quartiers de la ville.

[10] *L'unité urbaine* est une commune ou un ensemble de *communes* qui comporte sur son territoire une zone bâtie d'au moins 2 000 habitants où aucune habitation n'est séparée de la plus proche de plus de 200 mètres. En outre, chaque commune concernée possède plus de la moitié de sa population dans cette zone bâtie. Si l'unité urbaine s'étend sur plusieurs communes, l'ensemble de ces communes forme *une agglomération multicommunale* ou *agglomération urbaine*. Si l'unité urbaine s'étend sur une seule commune, elle est dénommée *ville isolée* (source : INSEE).

2. Revue de littérature

On retrouve habituellement les effets de diffusion (*spillover effect*) dans l'étude de l'interaction entre les marchés financiers en cas de crise financière ou dans l'étude du développement d'une nouvelle technologie. En recherche immobilière, Anselin, Florax et Rey (2004) indiquent que le terme « *spillover* » est utilisé quand la valeur d'une observation dépend des autres observations voisines. LeSage et Pace (2009) argumentent que le modèle autorégressif spatial (SAR) ou le modèle spatial de Durbin (SDM) sont les deux modèles qui prennent en compte cet effet de diffusion contrairement au modèle spatial des erreurs (SEM) qui n'est pas développé pour étudier l'effet de diffusion. Le modèle autorégressif spatial est largement utilisé pour étudier la dépendance des prix immobiliers (Brasington (1999); Can et Megbolugbe (1997); Can (1990); Can (1992); Pace et Gilley (1997); Won Kim, Phipps et Anselin (2003)). Can (1990) indique que le terme autorégressif ajouté au modèle hédonique permet de capturer l'effet de diffusion, et de mesurer l'influence du prix d'une maison sur les maisons voisines. Dans le modèle SAR, le paramètre ρ de la variable dépendante décalée mesure le degré d'influence des observations voisines sur l'observation en question. La valeur de ρ est comprise[11] entre 0 et 1. Une valeur positive élevée montre une forte dépendance entre les observations, tandis qu'une valeur de ρ proche de 0 indique que l'influence des observations voisines est minime. La revue de littérature détaillée sur l'utilisation du modèle autorégressif spatial en étude immobilière se trouve dans la partie 3.1 du CHAPITRE IV.

On a beaucoup parlé des effets de diffusion (*spillover effect)* après la crise des *subprimes*. Après la crise des *subprimes,* plusieurs études sur le marché immobilier américain se sont intéressées à l'impact négatif sur les prix de vente des biens voisins dans le cas de saisies immobilières. Lors d'une saisie d'un bien résidentiel, la « valeur à la casse[12] » est souvent inférieure au prix de marché. Cette baisse de prix se diffuse sur la valeur des autres biens voisins (*spillover effect*) en passant par le processus d'évaluation. Lin, Rosenblatt et Yao (2009) étudient les effets de diffusion (*spillover effect)* de la valeur

[11] Il n'est pas raisonnable d'envisager une valeur négative de ρ.

[12] La notion de « valeur à la casse » correspond à la valeur retenue par les créanciers dans le cas de saisie immobilière, cette valeur est souvent très éloignée de la valeur de marché.

à la casse d'un bien immobilier lors d'une procédure de saisie sur le prix des biens voisins et trouvent que, pour le marché immobilier à Chicago, cette baisse de prix crée une baisse de prix maximale de 8,7% sur les biens localisés dans un rayon de 0,9 kilomètres. Schuetz, Been et Ellen (2008) et Rogers et Winter (2009) trouvent le même phénomène sur le marché immobilier de New York et Saint Louis. Les deux études montrent que la baisse de valeur d'un bien entraine la baisse de valeur des biens voisins.

3. Données de transactions immobilières en France

L'étude du marché immobilier français se focalise habituellement sur le marché parisien. Cela peut être dû au manque de données ou parce que les investisseurs considèrent Paris comme la ville la plus représentative du marché immobilier français. Grâce à une base de données plus large, notre travail permet d'avoir une vision globale sur le marché immobilier français, de comparer le niveau de corrélation spatiale de différentes villes de France et, principalement, de vérifier si le marché parisien présente réellement un rôle dominant dans le marché immobilier français.

La base de données utilisée dans cette étude contient les valeurs de transactions sur des appartements dans les douze plus grandes agglomérations urbaines de France, à savoir Bordeaux, Lille, Lyon, Marseille, Montpellier, Nantes, Nice, Orléans, Paris, Rennes, Strasbourg et Toulouse. Chaque agglomération urbaine est composée de plusieurs communes. Le Tableau V.1 détaille le nombre de communes, la superficie et la taille de la population de ces douze agglomérations, selon les données de l'Insee publiées en 2010. Le nombre de transactions dans la base de données est 216 664, composé de 94 143 transactions pour 1998 et 122 521 transactions pour 2007. Les transactions sur les biens parisiens représentent la proportion la plus nombreuse parmi les douze agglomérations, soit 27 052 pour 1998 (donc 28,7%) et 28 828 pour 2007 (dont 23,5%). Orléans est la ville la plus petite avec le plus faible volume de transactions. Le volume important de transactions parisiennes, qui représente environ un quart du volume total de transactions chaque année, suggère que Paris aurait un rôle directeur dans l'évolution des prix des biens immobiliers français. Cependant, le volume de transactions à lui seul ne

suffirait pas à caractériser la ville dominante ; il faut vérifier si le mouvement des prix immobiliers en province suit celui de la capitale.

Le Tableau V.1 donne les prix moyens des transactions de chaque agglomération pour les différentes catégories des appartements : l'appartement standard, le studio (une pièce avec une salle de bain ou une salle d'eau), l'appartement standard d'une seule pièce, l'appartement de deux ou trois pièces et l'appartement de quatre pièces et plus. La valeur en parenthèse est le numéro d'ordre du niveau de prix moyen. La Figure V.1 compare les prix moyens des 15 villes de France (les 12 agglomérations et les 3 communes de la petite couronne de Paris) en 1998 et 2007. En 1998, le prix moyen des transactions immobilières dans la majorité des villes en France varie de 1100 €/m^2 à 1600 €/m^2, sauf à Paris où le prix moyen est environ 2400 €/m^2 et à Nanterre où le prix moyen est environ 2000€/m^2. Les prix au mètre carré en région Ile-de-France sont plus élevés qu'en province à l'exception de la commune de Seine-Saint-Denis avec ses caractéristiques particulières que sont le problème de délinquance et le taux d'immigration élevé. La communauté urbaine de Nice est une autre agglomération en province dont le niveau de prix moyen est élevé grâce à son attrait touristique. Les appartements dans cette région sont souvent des résidences secondaires appartenant aux étrangers. Le même phénomène est observé pour l'année 2007 mais avec des écarts des prix par agglomération plus élevés qu'en 1998.

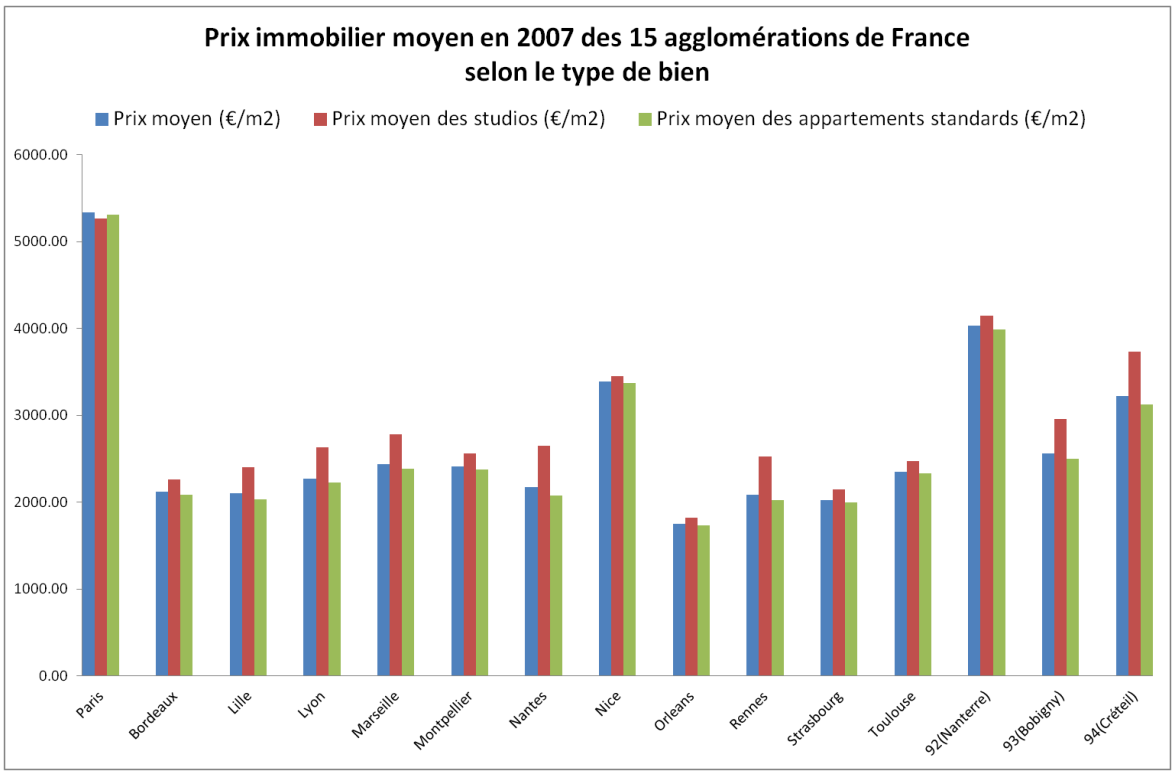
Tableau V.1 : Statistiques descriptives des valeurs des transactions immobilières des 12 agglomérations et des 3 communes de la petite couronne de Paris

	Paris	Bordeaux	Lille	Lyon	Marseille	Montpellier	Nantes	Nice	Orleans	Rennes	Strasbourg	Toulouse	92(Nanterre	93(Bobigny)	94(Créteil)
Nombre des communes	412	64	59	130	49	22	24	51	19	13	23	73			
Population	2125851	215374	184647	445274	797491	225511	270343	343123	113089	206194	263941	390301			
Superficie (km2)	105.4	49.4	30.2	47.9	240.6	56.9	65.2	71.9	27.5	50.4	78.3	118.3			
Range Nationale de la superficie	1	9	13	3	2	8	6	5	32	10	7	4			
Densité (population/km2)	20169.36	4359.80	6114.14	9295.91	3314.59	3963.29	4146.37	4772.23	4112.33	4091.15	3370.89	3299.25			
Range Nationale de la densité	1	5	3	2	11	9	6	4	7	8	10	12			
Nombre des transactions															
1998	27,052	2,929	2,127	9,498	6,583	2,336	3,262	10,379	1,653	1,907	2,461	3,651	9,857	4,544	5,904
2007	28,828	3,898	2,941	10,312	5,849	2,942	4,094	13,294	1,683	3,667	4,132	5,649	16,064	8,484	10,684
Total	55,880	6,827	5,068	19,810	12,432	5,278	7,356	23,673	3,336	5,574	6,593	9,300	25,921	13,028	16,588
Année 1998															
Prix moyen (€/m2)	2401.25 (1)	1063.75 (14)	1200.20 (9)	1147.13 (12)	1008.60 (15)	1223.08 (6)	1119.63 (13)	1517.12 (4)	1202.71 (8)	1188.05 (10)	1325.62 (5)	1162.76 (11)	2099.81 (2)	1209.60 (7)	1607.91 (3)
Prix moyen des studios (€/m2)	2286.40 (1)	1233.43 (15)	1425.82 (7)	1502.02 (4)	1270.50 (13)	1466.58 (6)	1316.17 (12)	1495.90 (5)	1379.80 (10)	1416.10 (9)	1418.26 (8)	1261.19 (14)	2149.54 (2)	1371.94 (11)	1671.24 (3)
Prix moyen des appartements standards (€/m2)	2416.83 (1)	1017.86 (14)	1149.17 (10)	1091.57 (12)	956.38 (15)	1154.57 (8)	1085.24 (13)	1520.56 (4)	1157.26 (7)	1149.44 (9)	1306.95 (5)	1137.48 (11)	2075.72 (2)	1181.69 (6)	1592.11 (3)
Prix moyen des biens à 1 pièce (€/m2)	2286.87 (1)	1225.13 (15)	1418.72 (7)	1462.32 (4)	1251.43 (13)	1458.81 (5)	1299.03 (11)	1495.29 (3)	1368.05 (10)	1418.97 (6)	1405.80 (8)	1255.03 (12)	2149.55 (2)	1372.03 (9)	1671.24 (3)
Prix moyen des biens à 2 ou 3 pièces (€/m2)	2313.80 (1)	1135.55 (13)	1205.39 (11)	1129.28 (14)	967.77 (15)	1269.05 (6)	1158.02 (12)	1536.03 (4)	1228.09 (8)	1208.85 (9)	1370.02 (5)	1241.36 (7)	1993.36 (2)	1206.05 (10)	1590.33 (3)
Prix moyen des biens à 4 pièces et plus (€/m2)	2823.31 (1)	778.36 (15)	1023.95 (9)	1058.54 (8)	956.40 (11)	893.80 (14)	951.39 (12)	1470.15 (4)	975.11 (10)	1062.72 (7)	1226.39 (5)	909.57 (13)	2291.06 (2)	1120.95 (6)	1618.79 (3)
Année 2007															
Prix moyen (€/m2)	5335.82 (1)	2116.39 (11)	2098.73 (12)	2272.71 (9)	2440.88 (6)	2409.31 (7)	2176.62 (10)	3387.71 (3)	1746.05 (15)	2080.75 (14)	2022.16 (13)	2351.23 (8)	4036.11 (2)	2561.40 (5)	3218.02 (4)
Prix moyen des studios (€/m2)	5268.59 (1)	2259.10 (13)	2400.72 (12)	2633.29 (8)	2784.94 (6)	2561.14 (9)	2651.31 (7)	3447.01 (4)	1819.45 (15)	2525.39 (10)	2147.39 (14)	2471.60 (11)	4144.25 (2)	2952.77 (5)	3729.60 (3)
Prix moyen des appartements standards (€/m2)	5307.70 (1)	2086.64 (10)	2032.46 (12)	2228.50 (9)	2384.23 (6)	2377.46 (7)	2078.44 (11)	3370.24 (3)	1731.40 (15)	2023.51 (13)	1994.76 (14)	2332.58 (8)	3986.56 (2)	2494.82 (5)	3128.10 (4)
Prix moyen des biens à 1 pièce (€/m2)	5269.79 (1)	2255.06 (13)	2390.11 (12)	2611.38 (8)	2773.08 (6)	2559.29 (9)	2643.88 (7)	3448.86 (4)	1818.66 (15)	2541.46 (10)	2152.01 (14)	2473.36 (11)	4145.29 (2)	2951.16 (5)	3728.45 (3)
Prix moyen des biens à 2 ou 3 pièces (€/m2)	5165.75 (1)	2201.15 (10)	2079.37 (13)	2302.73 (9)	2491.64 (6)	2491.51 (7)	2172.85 (11)	3403.22 (3)	1802.16 (15)	2089.41 (12)	2038.24 (14)	2424.10 (8)	3913.80 (2)	2599.18 (5)	3201.26 (4)
Prix moyen des biens à 4 pièces et plus (€/m2)	5919.33 (1)	1803.68 (14)	1903.18 (12)	2129.27 (7)	2235.63 (5)	2020.83 (8)	1911.86 (11)	3245.88 (3)	1585.54 (15)	1893.51 (13)	1958.93 (10)	2004.35 (9)	4234.88 (2)	2233.59 (6)	2995.19 (4)
Distance de Paris (kilomètre)	0	499.28	204.69	392.25	661.36	595.30	342.14	687.30	111.06	308.24	396.96	588.08	11.25	9.32	12.12

Le numéro entre parenthèses est le numéro d'ordre (du plus petit au plus grand)

Figure V.1 : Prix moyens de biens immobiliers de 12 agglomérations et 3 communes de la petite couronne de Paris selon le type de bien en 2007 (A) et en 1998 (B)

(A)



(B)

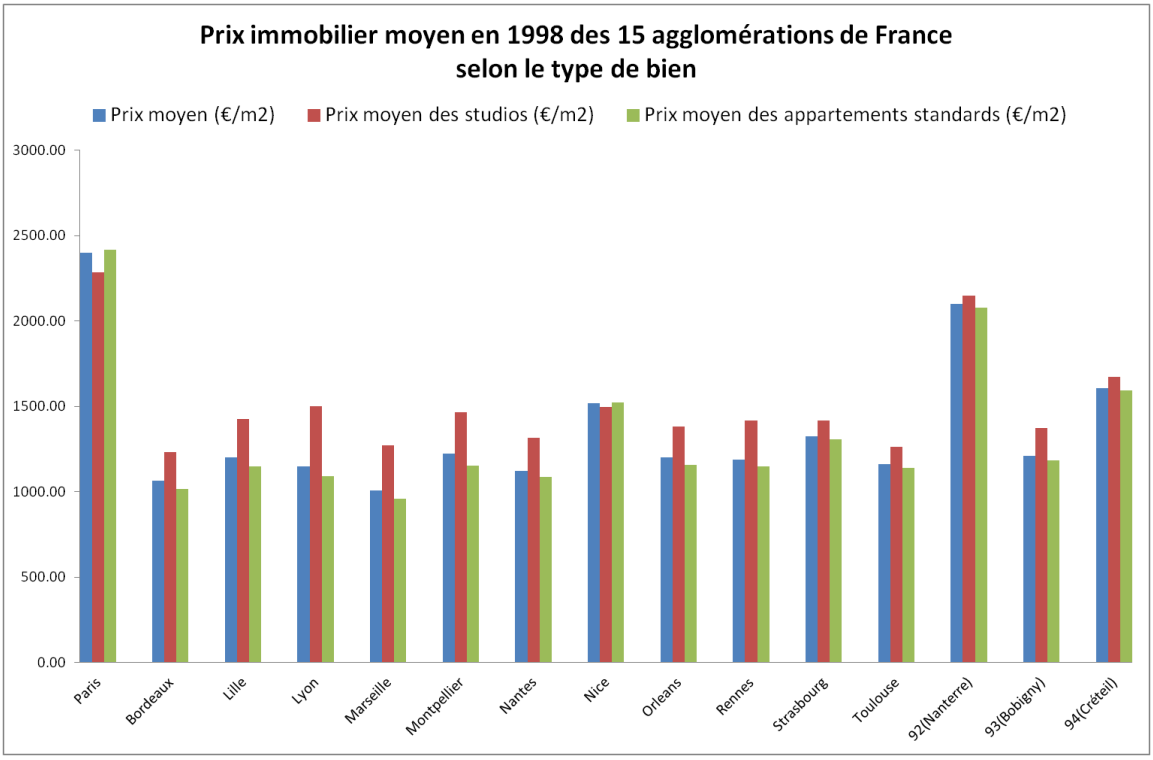
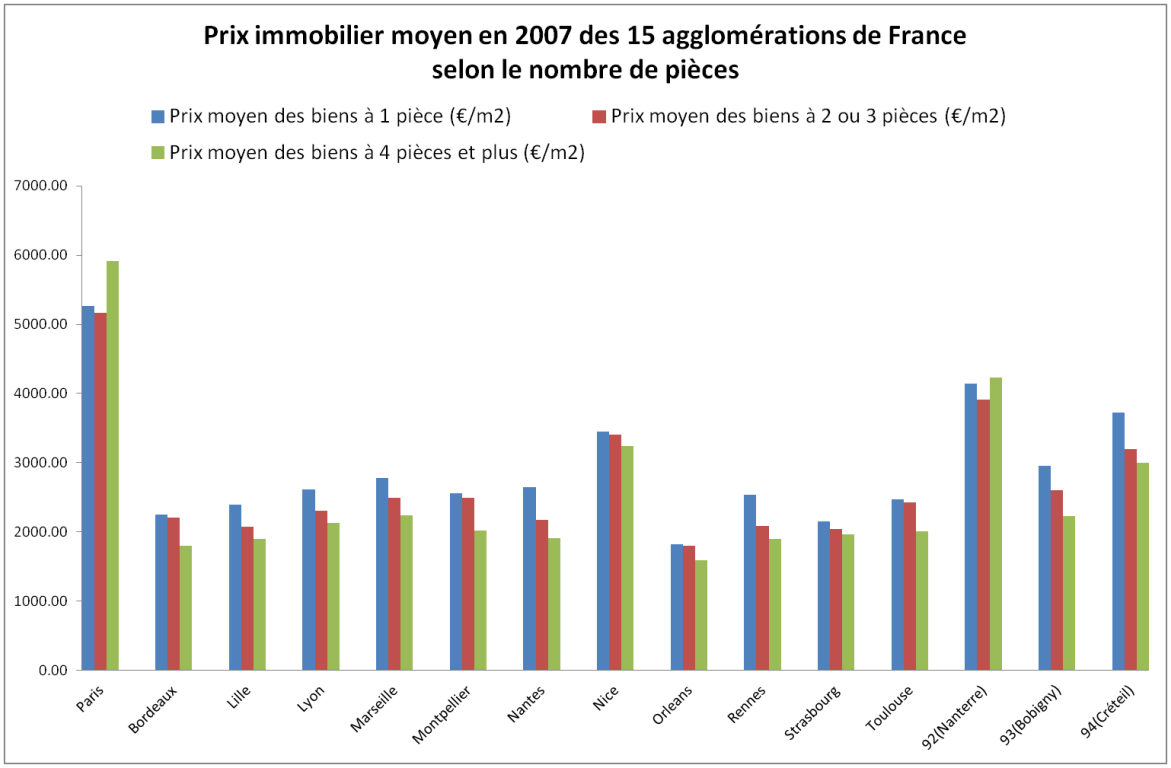
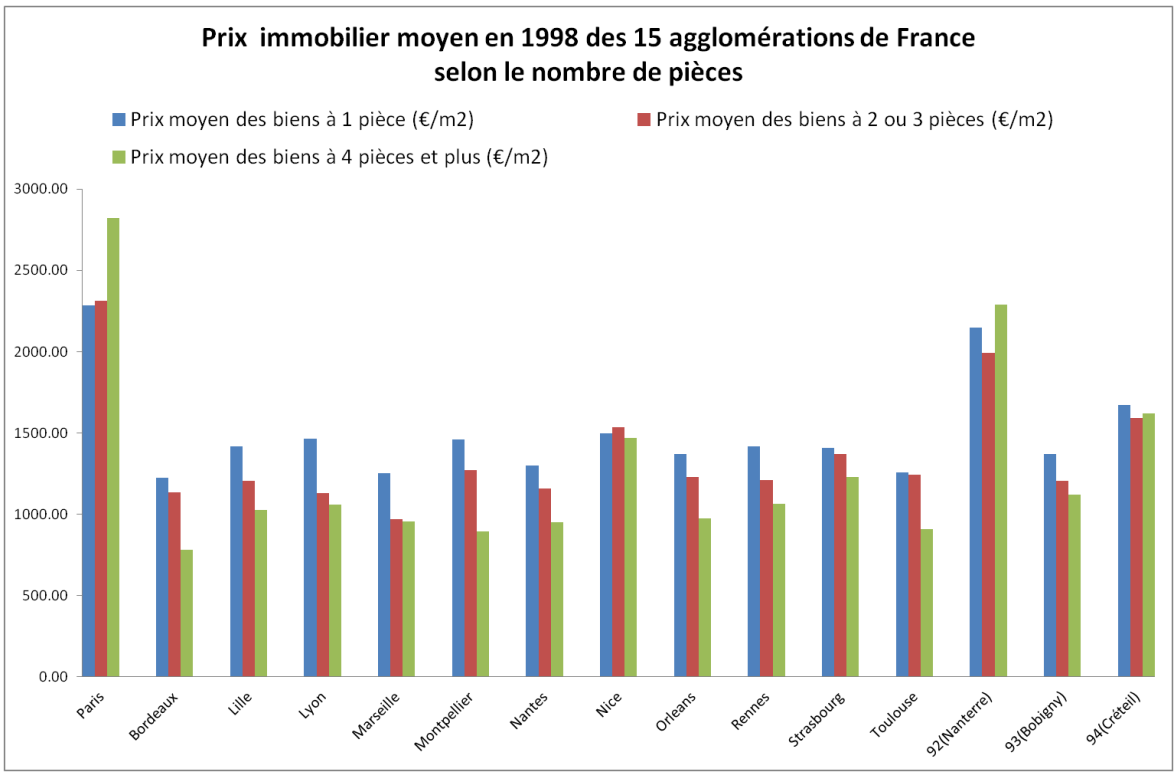


Figure V.2 : Prix moyens de biens immobiliers de 12 agglomérations et 3 communes de la petite couronne de Paris selon le nombre de pièce en 2007 (A) et en 1998 (B)
(A)



(B)



Observons maintenant le prix moyen par type de bien (Figure V.1). Pour toutes les agglomérations urbaines hors Paris, le prix de vente (au mètre carré) du studio est souvent un peu plus cher que celui d'un logement standard. Ces résultats correspondent bien à la réalité du marché immobilier. Cette différence de prix peut s'expliquer par plusieurs facteurs tels que la localisation des studios dans le centre ville, le mode de vie des jeunes générations et étudiants. A contrario, le prix du studio à Paris est un peu plus faible que celui d'un appartement standard. Ce résultat qui paraît différent à ce que nous observons en réalité peut s'expliquer par la diversité des biens inclus dans la catégorie studio. A Paris, il existe deux types de studio : le studio standard et les chambres de bonnes transformés en studio. Ces chambres de bonnes se trouvent habituellement au dernier étage sans ascenseur, et leur prix est inférieur à celui d'un studio standard ; ce qui explique cette baisse du prix moyen comparé à un appartement standard à Paris.

En rapportant le prix moyen au nombre de pièces (Figure V.2), nous observons que pour les 15 villes étudiées hors Paris et Nanterre, le prix au mètre carré d'un appartement de 2 ou 3 pièces est supérieur à celui d'appartement de 4 pièces et plus.

4. Méthodologie

Ce travail se divise en deux niveaux : l'analyse au niveau agrégé et l'analyse au niveau des transactions. L'analyse au niveau agrégé donne une vision globale de l'autocorrélation des valeurs immobilières françaises et permet de vérifier le rôle de Paris sur le marché immobilier français. Le deuxième niveau d'analyse est un niveau plus détaillé qui étudie l'autocorrélation spatiale entre les prix de transactions. L'objectif de ce niveau d'analyse est de déterminer s'il existe un quartier (ou un arrondissement) qui présenterait un rôle dominant sur le marché immobilier de chaque ville. Les détails de ces deux niveaux d'analyse et les hypothèses associés forment les deux sections suivantes.

4.1. Etude du niveau agrégé

Ce premier niveau d'analyse a pour l'objectif de mesurer le degré de corrélation spatiale entre les différentes communes de France et d'étudier le rôle du marché immobilier parisien sur l'ensemble du marché immobilier français en comparant les degrés de corrélation spatiale obtenus à partir des bases des données avec et sans les données parisiennes. Les questions posées à ce niveau d'analyse sont : existe-t-il un lien entre les prix immobiliers à Paris et ceux de la province ? Paris représente-t-elle comme une ville dominante sur le marché immobilier français ? Pour répondre à ces questions, le degré de corrélation spatiale sera estimé, pour la première étape, avec l'ensemble de la base de données. Ensuite, le paramètre de corrélation spatiale sera ré-estimé avec les mêmes données hors celles de Paris et celles des trois petites couronnes d'Ile-de-France. Si, en l'absence des données de Paris, le degré de la dépendance spatiale se réduit considérablement, alors les prix immobiliers des villes en province sont fortement corrélés avec les prix immobiliers parisiens. Cela confirmerait alors que Paris est une ville dominante dans le marché immobilier en France.

Ce travail utilise les données au niveau agrégé. Une unité d'étude correspond à l'arrondissement pour les 3 grandes villes de France (20 arrondissements de Paris, 16 arrondissements de Marseille, 9 arrondissements de Lyon) et les communes des 12 agglomérations avec plus de 200 transactions. Le modèle utilisé pour étudier l'autocorrélation spatiale est le modèle autorégressif spatial (SAR) qui prend en compte l'autocorrélation spatiale des variables endogènes. L'équation de régression du niveau macro est donc :

$$\bar{y} = \alpha i_n + \rho W_{SAR}\bar{y} + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2 I_n) \qquad \text{Eq. V.1}$$

$\bar{y}$ est le prix moyen de chaque commune, αi_n est le vecteur de la constante. Le paramètre ρ représente le degré de dépendance spatiale, ce paramètre servant à mesurer l'influence des prix moyen des communes voisines sur la commune en question. La matrice W_{SAR} est

la matrice de poids correspondant au modèle SAR. Plusieurs types de matrice de poids ont été choisis pour cette analyse, de sorte à pouvoir étudier la sensibilité du résultat par rapport au type de matrice de poids. Les matrices de poids utilisées sont les suivantes :

- la matrice de voisinage avec la contiguïté d'ordre 1 (le voisin le plus proche) et la pondération standardisée,
- la matrice de voisinage avec la contiguïté d'ordre 1 (le voisin le plus proche) et l'inverse de la distance comme pondération,
- la matrice de voisinage avec la contiguïté d'ordre 4 et la pondération standardisée,
- la matrice de voisinage avec la contiguïté d'ordre 4 et l'inverse de la distance comme pondération,
- la matrice de voisinage avec la condition de distance séparation maximale de 150 mètres et la pondération standardisée,
- la matrice de voisinage avec la condition de distance séparation maximale de 150 mètres et l'inverse de la distance comme pondération.

Toutes les étapes de ce travail ont recours au logiciel *R* avec le package *spdep* qui fournit plusieurs fonctions permettant d'estimer le degré de corrélation spatiale. Les fonctions utilisées sont les suivantes :

- *knearneigh*: la fonction retourne la matrice de voisinage selon la contiguïté d'ordre *k*.
- *dnearneigh*: la fonction identifie les voisins de chaque observation selon la condition de distance fixée. L'option *longlat = TRUE* permet d'utiliser la distance à vol d'oiseau en kilomètres à la place de la distance euclidienne.
- *nb2listw* : la fonction crée la matrice de poids à partir de la matrice de voisinage. Plusieurs options sont proposées, *W* pour le poids standardisé, *B* pour le poids binaire, *U* pour le poids binaire avec la somme de chaque ligne égale à 1 ou *glist*=.. pour fixer la condition du calcul de poids (l'inverse de la distance ou l'inverse de la distance au carrée).

- *listw2U* : la fonction rend symétrique la matrice de poids. Cette fonction est nécessaire pour rend symétrique la matrice de contiguïté (qui est parfois asymétrique).
- *moran.test* : la fonction permet de vérifier la présence de corrélation spatiale pour chaque variable fixée.
- *lagsarlm* : la fonction permet de faire une régression linéaire avec la présence de corrélation spatiale des variables endogènes (modèle SAR).

Remarquons que notre travail prend uniquement en compte l'autocorrélation spatiale des variables endogènes (modèle SAR). Comme l'objectif principal de ce travail est de déterminer un quartier dominant ou une ville dominante du marché immobilier français, la source de l'autocorrélation spatiale choisie est le processus d'évaluation. Ce processus d'évaluation par le vendeur ou par l'expert immobilier crée un effet de diffusion (*spillover effect*). Dans ce travail, nous supposerons que la valeur d'un bien immobilier dépend uniquement de la valeur des autres biens voisins, de sorte que la ressemblance des biens voisins ou l'influence de l'externalité ne sont pas prises en compte. Par ailleurs, le test du multiplicateur de Lagrange[13] confirme ce choix en donnant un résultat significatif pour le paramètre de corrélation du modèle SAR.

A ce niveau d'analyse, le degré de corrélation ρ est estimé en combinant les conditions suivantes :

- les 3 bases des données : toutes les communes (ρ_{toutes}), toutes les communes sauf Paris ($\rho_{sauf\ Paris}$) et toutes les communes sauf les communes en Ile-de-France ($\rho_{sauf\ IdF}$),
- les 2 années d'étude : 1998 et 2007,
- les 2 conditions de voisinage : la contiguïté et la condition de distance de 150[14] kilomètres ou 444 kilomètres,
- les 2 conditions des poids : les poids standardisés et l'inverse de la distance

[13] Ce test permet à vérifier la présence de dépendance spatiale dans le modèle de régression linéaire. Le résultat du multiplicateur de Lagrange donne une p-value < 0,01% pour le modèle SAR.

[14] La distance maximum fixée pour la condition de distance est 150 kilomètres pour l'année 2007 et 444 kilomètres pour l'année 1998. Ces deux valeurs sont choisies pour que chaque observation contienne au moins un voisin.

- les 6 prix moyens des différentes types de biens : les prix de tous les types de biens, les prix moyens des studios uniquement, les prix moyens des appartements standards uniquement, les prix moyens des appartements avec une seule pièce uniquement, les prix moyens des appartements avec deux ou trois pièces uniquement et les prix moyens des appartements avec plus de trois pièces uniquement.

Nos hypothèses de travail sont les suivantes :

Pour chaque cas d'étude, ρ_{toutes} est comparé à $\rho_{sauf\ Paris}$ et $\rho_{sauf\ IdF}$.
Si $\rho_{sauf\ Paris} < \rho_{toutes}$, alors Paris présente un rôle dominant dans le marché immobilier français. Cette baisse du niveau de corrélation spatiale signifie que Paris présente un rôle directif dans l'effet de diffusion et les prix immobiliers des autres quartiers sont partiellement corrélés avec le prix immobilier de cette ville dominant.
Si $\rho_{sauf\ IdF} < \rho_{sauf\ Paris} < \rho_{toutes}$, alors la région Ile-de-France est une région dominante du marché immobilier français.

4.2. Etude du niveau des transactions

L'étude se poursuit au niveau des transactions avec des données plus détaillées. Ce niveau d'analyse s'intéresse davantage à l'autocorrélation à l'intérieur de chaque commune. Les données utilisées sont les valeurs des transactions réelles. Le volume élevé de transactions permet de vérifier de façon plus précise quelle est la structure de la corrélation spatiale dans chaque commune. L'objectif de cette analyse est d'estimer le degré de corrélation spatiale à l'intérieur de chaque ville et de déterminer si pour chaque ville, s'il existe une zone dominante. Les questions posées à cette étape sont les suivantes : le centre ville présente-t-il une zone dominante pour la détermination des prix dans les autres communes regroupées dans la même unité urbaine ? Comment peut-on confirmer économétriquement que le centre ville présente le quartier dominant de marché immobilier de la ville ? Et dans le cas d'une ville polycentrique, comment peut-on déterminer le quartier dominant ?

Les données utilisées dans ce travail sont les valeurs des transactions dans douze agglomérations (Bordeaux, Lille, Lyon, Marseille, Montpellier, Nantes, Nice, Orléans, Paris, Rennes, Strasbourg et Toulouse). Le paramètre de corrélation spatiale est dans un premier temps estimé à partir des transactions de toutes les villes afin de déterminer le degré de corrélation spatiale à l'intérieur de la ville. Ce paramètre est ensuite ré-estimé avec les mêmes données mais en excluant tour à tour les données de chaque arrondissement (pour les trois villes ; Paris, Lyon et Marseille) ou celles du centre ville (pour les neuf autres villes).

A ce niveau d'étude les caractéristiques physiques des biens sont incluses dans le modèle de régression, cela permet d'améliorer le pouvoir explicatif du modèle. L'équation de régression spatiale selon le modèle autorégressif spatial (SAR) est présentée par les équations suivantes :

$$y = \alpha\iota_n + X\beta + \rho W_{SAR} y + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2 I_n) \qquad \text{Eq. V.2}$$

Comme à l'Eq. V.2, X représente une matrice des variables explicatives (variables exogènes) de dimension $n \times k$ et β est le vecteur $k \times 1$ des coefficients liés aux caractéristiques non spatiales prises en compte dans cette régression. Les éléments de la matrice X sont les caractéristiques physiques du bien et les caractéristiques socio-dermographiques de l'acheteur. Les dix caractéristiques physiques prises en compte dans cette étude sont les suivantes :

- *NbRoom* (Nombre des pièces), mesuré par 5 variables dummies; *NbRoom1*, *NbRoom2*, *NbRoom3*, *NbRoom4* et *NbRoom5plus* correspond au nombre de pièces de 1, 2, 3, 4 ou 5 et plus. *NbRoom2* est défini comme la catégorie de référence.
- *NbBath* (Nombre de salles de bien ou salles d'eau), mesuré par 3 variables dummies; *NbBath1*, *NbBath2* et *NbBath3plus* correspond au nombre de salles de bain de 1, 2 ou 3 et plus. *NbBath1* est défini comme la catégorie de référence.

- *Floor* (Etage), mesuré par 8 variables dummies; *Floor0, Floor1, Floor2, Floor3, Floor4, Floor5, Floor6, Floor7plus* correspond à l'appartement situé au rez-de-chaussée, au 1er, 2ème, 3ème, 4ème, 5ème, 6ème ou 7ème étage et plus. *Floor2* est défini comme la catégorie de référence.
- *AprtTyp* (Type de l'appartement), mesuré par 5 variables dummies; *AS* (Appartement Standard), *DU* (Appartement Duplex) et *ST* (Studio). *AS* est défini comme la catégorie de référence.
- *Period* (Période de construction), mesuré par 5 variables dummies; *Na* (non enseignée), *Bf1913* (Avant 1913), *1914_1947*, *1948_1980*, *Af1980* (Après 1980). *Na* est défini comme la catégorie de référence.
- *Elevator* (Existence de l'ascenseur), la variable du type dummy permettant la prise en compte de l'existence de l'ascenseur (Oui/Non) dans le modèle. *Elev=N* est défini comme la catégorie de référence.
- *Parking* (Existence du garage), la variable du type dummy permettant la prise en compte de l'existence du garage (Oui/Non) dans le modèle. *Parking* = 0 est défini comme la catégorie de référence.
- *ExtraRoom* (Existence d'une chambre de service), la variable du type dummy permettant la prise en compte de l'existence d'une chambre de service (Oui/Non) dans le modèle. *ExtraRoom* = 0 est défini comme la catégorie de référence.
- *Terrace* (Existence d'une terrasse), la variable du type dummy permettant la prise en compte de l'existence d'une terrasse (Oui/Non) dans le modèle. *Terrace* = 0 est défini comme la catégorie de référence.
- *Garden* (Existence du jardin), la variable du type dummy permettant la prise en compte de l'existence du jardin (Oui/Non) dans le modèle. *Garden* = 0 est défini comme la catégorie de référence.

Les quatre caractéristiques socio-dermographiques de l'acheteur qui sont prises en compte sont les suivantes :

- *a_prof* (Profession de l'acheteur), mesuré par 7 variables dummies; *aprof=na* (non enseigné), *a_prof=12* (Agriculteur, artisan et commerçant)

a_prof=3 (Profession libéral, Professeur et Cadre), *a_prof=4* (Profession intermédiaire et technicien) *a_prof=56* (Employeur et ouvrier) *a_prof=7* (Retraité), *a_prof=8* (Sans activité professionnelle). *a_prof=na* est défini comme la catégorie de référence.

- *a_matri* (Statut matrimoniale de l'acheteur), mesuré par 3 variables dummies; *C* (Célibataire), *M* (Marié, remarié ou pacsé), *V* (Veuf ou divorcé). *a_matri=C* est défini comme la catégorie de référence.
- *a_sexe* (Sexe de l'acheteur), mesuré par 2 variables dummies; *M* (Masculine) et *F* (Féminine). *a_sexe=M* est défini comme la catégorie de référence.
- *a_age* (Age de l'acheteur), mesuré par 3 variables dummies; *<30* (Moins de 30 ans), *30-59* et *>60* (Plus de 60 ans). *a_age=30-59* est défini comme la catégorie de référence.

Le paramètre ρ représente le paramètre de la dépendance spatiale $(0 < \rho < 1)$. La matrice de poids, W_{SAR}, applique la condition de distance comme la condition de voisinage avec le poids correspond à l'inverse de la distance. Les éléments de cette matrice de poids, w_{ij}, sont définis de la façon suivante :

$$w_{ij} = \begin{cases} \frac{1}{d_{ij}}, & d_{ij} < d^* \\ 0, & d_{ij} \geq d^* \end{cases} \qquad \text{Eq. V.3}$$

$$i, j = 1, \dots, n$$

n est le nombre des observations, d_{ij} est la distance séparant les biens *i* et *j*, d^* est la distance seuil, pondérée par la densité de population de chaque ville étudiée. Si la distance entre deux biens est inférieure à la distance seuil, alors ces deux biens sont considérés comme des biens voisins et l'élément associé dans la matrice de poids est égale à l'inverse de cette distance, et à 0 sinon.

5. Résultats et interprétation

5.1. Autocorrélation spatiale des prix immobiliers en France

Le Tableau V.2 présente les valeurs de $\hat{\rho}$ estimées de l'Eq. V.1 avec les données des communes de France pour l'année 2007 et le Tableau V.3 présente les mêmes résultats pour l'année 1998. Le degré de dépendance spatiale est estimé à partir de trois groupes de données (toutes les données, toutes les données sauf Paris et toutes les données sauf les communes d'Ile-de-France), pour les deux années d'étude (1998 et 2007), pour les trois conditions de voisinage (contiguïté d'ordre 1, contiguïté d'ordre 4 et condition de distance de 150km [15] ou 444km), pour les 2 conditions des poids (standardisé et inverse de la distance) et pour les prix moyens des six différents types de biens (les prix de tous les types, les prix moyens des studios uniquement, les prix moyens des appartements standards uniquement, les prix moyens des appartements avec une seule pièce uniquement, les prix moyens des appartements avec deux ou trois pièces uniquement et les prix moyens des appartements avec plus de trois pièces uniquement).

Les $\hat{\rho}$ sont, en majorité, significatifs. Les corrélations estimées avec l'ensemble des prix moyens des 141 communes en 2007 sont compris entre 0,5932 à 0,8374. Ces résultats montrent une forte corrélation spatiale des prix immobiliers des 141 communes de France en 2007. Cette corrélation varie selon la condition de voisinage, la condition de poids et le type du bien. Le degré de corrélation estimé avec la contiguïté d'ordre 1 et les pondérations standardisés (Tableau V.2 – A) est de 0,6199 et cette valeur s'élève à 0,8358 si la condition de voisinage est la contiguïté d'ordre 4. Ce résultat paraît logique car la contiguïté d'ordre 4 permet de prendre en compte un nombre de voisinages plus élevé que la contiguïté d'ordre 1. Cette hausse de valeur de $\hat{\rho}$ signifie que les prix immobiliers d'une commune sont corrélés non pas uniquement aux prix immobiliers de la commune la plus proche mais aussi aux prix des communes aux alentours. L'autocorrélation est égale

[15] La distance maximum de la condition de voisinage est fixée à 150 kilomètres pour l'année 2007 pour que chaque observation ait au moins un voisin. Pour cette même raison, la distance maximum est fixée à 444 kilomètres pour l'année 1998.

à 0,7071 si les voisinages sont déterminés avec la condition de distance seuil de 150 kilomètres. L'utilisation de la condition de distance pour définir le voisinage permet à prendre uniquement en compte le voisinage très proche. Mais si la distance seuil choisie est trop faible, il peut y avoir des observations isolées. C'est la raison pour laquelle nous avons fixé la distance seuil à 150 kilomètres (en 1998) et à 444 kilomètres (en 2007), ces distances seuils correspondent à la distance minimale pour laquelle chaque observation a au moins un voisin.

Les estimations de $\hat{\rho}$ varient aussi avec la condition de pondération. Le Tableau V.2 – B diffère du Tableau V.2 – A par la pondération choisie. Le Tableau V.2 – A applique la pondération standardisée où à chaque voisin est attribué le même poids égal à l'inverse du nombre total des voisins. Pour le Tableau V.2 – B, les poids attribués sont l'inverse de la distance. La pondération par l'inverse de la distance est plus souvent choisie dans le cas de l'analyse de l'effet de diffusion car l'influence des observations voisines sur une observation est de plus en plus faible au fur et à mesure que l'on s'éloigne de cette observation. Le niveau de l'influence varie inversement avec la distance entre les observations. Par contre, la pondération standardisée accorde la même importance à tous les voisins. En comparant le résultat du Tableau V.2 – A et Tableau V.2 – B, nous observons que le degré de corrélation spatiale varie faiblement si la pondération par l'inverse de la distance est remplacée la condition standardisée. Les degrés de corrélation trouvés sont 0,6199 ; 0,8039 et 0,7946 pour la contiguïté d'ordre 1, la contiguïté d'ordre 4 et la condition de distance seuil de 150 kilomètres.

Tableau V.2 : Degrés de corrélation estimés ($\hat{\rho}$) pour l'année 2007 avec les données du niveau macro, les différents prix et les différentes matrices de poids

Tableau (A)

Base de données	Condition de voisinage	Condition de poids	ρ (Prix07)	ρ (Prix07ST)	ρ (Prix07AS)	ρ (Prix07r1)	ρ (Prix07r23)	ρ (Prix07r4)
Données 2007	Contiguïté de 1er ordre	Standardisé	61.99% ***	59.36% ***	61.98% ***	59.32% ***	62.03% ***	61.52% ***
Données 2007 sauf Paris	Contiguïté de 1er ordre	Standardisé	54.17% ***	51.30% ***	54.23% ***	51.27% ***	54.15% ***	53.01% ***
Données 2007 Sauf Ile-de-France	Contiguïté de 1er ordre	Standardisé	53.67% ***	37.36% ***	54.03% ***	37.11% ***	56.32% ***	45.78% ***
Données 2007	Contiguïté de 4ème ordre	Standardisé	83.58% ***	80.08% ***	83.74% ***	80.07% ***	83.27% ***	83.91% ***
Données 2007 sauf Paris	Contiguïté de 4ème ordre	Standardisé	74.71% ***	70.84% ***	74.79% ***	70.95% ***	73.99% ***	75.03% ***
Données 2007 Sauf Ile-de-France	Contiguïté de 4ème ordre	Standardisé	67.99% ***	56.58% ***	69.04% ***	56.73% ***	68.86% ***	66.20% ***
Données 2007	Distance de 150 km	Standardisé	70.71% ***	72.61% ***	70.39% ***	72.77% ***	70.65% ***	68.76% ***
Données 2007 sauf Paris	Distance de 150 km	Standardisé	69.02% ***	70.00% ***	68.78% ***	70.21% ***	69.61% ***	65.32% ***
Données 2007 Sauf Ile-de-France	Distance de 150 km	Standardisé	6.40% *	5.92% *	6.50% *	5.92% *	6.18% *	7.02% *

Tableau (B)

Données 2007	Contiguïté de 1er ordre	Inverse de distance	61.99% ***	59.36% ***	61.98% ***	59.32% ***	62.03% ***	61.52% ***
Données 2007 sauf Paris	Contiguïté de 1er ordre	Inverse de distance	54.17% ***	51.30% ***	54.23% ***	51.27% ***	54.15% ***	53.01% ***
Données 2007 Sauf Ile-de-France	Contiguïté de 1er ordre	Inverse de distance	53.67% ***	37.36% ***	54.03% ***	37.11% ***	56.32% ***	45.78% ***
Données 2007	Contiguïté de 4ème ordre	Inverse de distance	80.39% ***	76.17% ***	80.62% ***	76.12% ***	80.11% ***	80.81% ***
Données 2007 sauf Paris	Contiguïté de 4ème ordre	Inverse de distance	70.63% ***	66.12% ***	70.80% ***	66.17% ***	70.11% ***	70.80% ***
Données 2007 Sauf Ile-de-France	Contiguïté de 4ème ordre	Inverse de distance	63.06% ***	49.36% ***	64.20% ***	49.38% ***	64.81% ***	58.84% ***
Données 2007	Distance de 150 km	Inverse de distance	79.46% ***	77.41% ***	79.55% ***	77.40% ***	79.47% ***	79.08% ***
Données 2007 sauf Paris	Distance de 150 km	Inverse de distance	75.84% ***	72.32% ***	76.02% ***	72.35% ***	76.08% ***	74.68% ***
Données 2007 Sauf Ile-de-France	Distance de 150 km	Inverse de distance	6.56% *	6.17% *	6.64% *	6.18% *	6.33% *	7.13% *

Les symboles *, ** et *** désignent la significativité à 5%, 1% et 0,01%, respectivement.

Tableau V.3 : Degrés de corrélation estimés ($\hat{\rho}$) pour l'année 1998 avec les données du niveau macro, les différents prix et les différentes matrices de poids

Tableau (A)

Base de données	Condition de voisinage	Condition de poids	ρ (Prix98)	ρ (Prix98ST)	ρ (Prix98AS)	ρ (Prix98r1)	ρ (Prix98r23)	ρ (Prix98r4)
Données 1998	Contiguïté de 1er ordre	Standardisé	58.42% ***	47.88% ***	62.58% ***	49.25% ***	61.61% ***	63.86% ***
Données 1998 sauf Paris	Contiguïté de 1er ordre	Standardisé	50.79% ***	38.66% ***	57.39% ***	40.37% ***	56.57% ***	56.92% ***
Données 1998 Sauf Ile-de-France	Contiguïté de 1er ordre	Standardisé	41.38% ***	22.73% *	55.13% ***	24.55% ***	54.36% ***	53.39% ***
Données 1998	Contiguïté de 4ème ordre	Standardisé	74.83% ***	68.18% ***	76.27% ***	69.69% ***	74.91% ***	77.97% ***
Données 1998 sauf Paris	Contiguïté de 4ème ordre	Standardisé	66.02% ***	58.97% ***	68.45% ***	61.06% ***	66.95% ***	69.99% ***
Données 1998 Sauf Ile-de-France	Contiguïté de 4ème ordre	Standardisé	56.22% ***	43.33% ***	61.69% ***	46.67% ***	60.21% ***	63.21% ***
Données 1998	Distance de 444 km	Standardisé	89.33% ***	88.19% ***	89.48% ***	87.87% ***	88.11% ***	91.61% ***
Données 1998 sauf Paris	Distance de 444 km	Standardisé	82.63% ***	80.88% ***	82.66% ***	81.00% ***	80.02% ***	86.84% ***
Données 1998 Sauf Ile-de-France	Distance de 444 km	Standardisé	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Tableau (B)

Données 1998	Contiguïté de 1er ordre	Inverse de distance	58.42% ***	47.88% ***	62.58% ***	49.25% ***	61.61% ***	63.86% ***
Données 1998 sauf Paris	Contiguïté de 1er ordre	Inverse de distance	50.79% ***	38.66% ***	57.39% ***	40.37% ***	56.57% ***	56.92% ***
Données 1998 Sauf Ile-de-France	Contiguïté de 1er ordre	Inverse de distance	41.38% ***	22.73% *	55.13% ***	24.55% ***	54.36% ***	53.39% ***
Données 1998	Contiguïté de 4ème ordre	Inverse de distance	73.85% ***	66.13% ***	75.53% ***	67.84% ***	74.14% ***	76.97% ***
Données 1998 sauf Paris	Contiguïté de 4ème ordre	Inverse de distance	64.35% ***	56.15% ***	67.06% ***	58.44% ***	65.61% ***	68.19% ***
Données 1998 Sauf Ile-de-France	Contiguïté de 4ème ordre	Inverse de distance	54.29% ***	39.62% ***	60.48% ***	43.11% ***	59.17% ***	61.01% ***
Données 1998	Distance de 444 km	Inverse de distance	82.25% ***	78.41% ***	83.13% ***	78.93% ***	82.43% ***	83.56% ***
Données 1998 sauf Paris	Distance de 444 km	Inverse de distance	77.94% ***	71.66% ***	79.66% ***	72.59% ***	78.89% ***	79.54% ***
Données 1998 Sauf Ile-de-France	Distance de 444 km	Inverse de distance	71.10% ***	54.81% ***	75.77% ***	56.07% ***	74.75% ***	74.50% ***

Les symboles *, ** et *** désignent la significativité à 5%, 1% et 0,01%, respectivement.

Les estimations de $\hat{\rho}$ varient avec les conditions de voisinage et la pondération choisies, mais toutes les valeurs de $\hat{\rho}$ confirment une forte corrélation spatiale des prix immobiliers des différentes communes en France en 2007. Ces mêmes résultats sont observés pour l'année 1998 mais avec un niveau d'autocorrélation spatiale un peu plus faible, sauf avec la condition de distance limite de 444 kilomètre où l'autocorrélation est nettement supérieure. Ce niveau très élevé de corrélation est dû à la distance seuil fixée qui est environ trois fois plus grande que celle de 2007. A partir de nos résultats de 1998 et 2007, nous établissons donc une forte corrélation spatiale dans les prix immobiliers des différentes communes en France.

En regardant la variation du degré de corrélation en fonction du type de bien, les prix de studios et d'appartements d'une seule pièce ont un niveau de corrélation plus faible comparé aux autres catégories. Cela peut s'expliquer par la spécificité de la demande de ces biens. Ces deux types de biens sont plus demandés dans les grandes villes qui comptent beaucoup d'étudiants et moins demandés dans les petites villes, ce qui crée donc une différence de prix pour ce type de biens.

5.2. Paris a-t-il un rôle directif dans le marché immobilier français ?

Pour pouvoir répondre à la question de savoir si *Paris est une ville dominante du marché immobilier français, le* degré de corrélation spatiale est ré-estimé sans les données de Paris.

Les résultats de l'année 2007 (Tableau V.2-B) montrent que, sans les données de Paris, la valeur ρ̂ baisse de 0,6199 à 0,5417. L'estimation de $\hat{\rho}$ baisse encore plus à 0,5367 si les données des communes en Ile-de-France ont été ignorées. Ce résultat persiste même si la condition de voisinage est changée, $\hat{\rho}$ baissant de 0,8039 à 0,6306 si la condition de voisinage est la contiguïté d'ordre 4. La baisse est encore plus forte si le voisinage est déterminé par la condition de distance : le degré de corrélation baisse de 0,7946 à 0,0656. Nous pouvons dire que le degré de corrélation spatiale est très faible si les données d'Ile-de-France sont exclues des données analysées. Cette baisse de $\hat{\rho}$ est

aussi observée pour l'année 1998 (Tableau V.3-B), où nous observons une baisse de 0,5842 à 0,5079 si les données de Paris sont exclues et une baisse à 0,4138 si les données d'Ile-de-France sont exclues. De plus, dans le cas de l'année 1998, si le voisinage est déterminé par la condition de distance avec la pondération standardisée (Tableau V.3-A), le paramètre de corrélation spatiale estimé n'est pas significatif. Cela indique bien qu'il n'existe pas de dépendance spatiale des prix parmi les biens immobiliers en province en 1998.

Pour ces deux années, le degré d'autocorrélation spatiale estimé sans les données de Paris paraît moins élevé que celui obtenu avec les données de Paris. A partir de ces résultats, nous arrivons donc à confirmer le rôle dominant de Paris sur le marché immobilier français. Les prix des biens immobiliers localisés dans les différentes communes en dehors d'Ile-de-France sont très faiblement corrélés mais les prix des biens en province sont corrélés aux prix immobiliers parisiens.

L'analyse se poursuit en regardant la sensibilité de ces résultats aux types de bien. Nous vérifions si ce résultat persiste pour tous les types de bien. Pour les appartements standards, une fois enlevées les données d'Ile-de-France, le degré de corrélation baisse de 0,6198 à 0,5403 pour l'année 2007 et de 0,6258 à 0,5513 pour l'année 1998. Comparé aux résultats estimés à partir de tous les types de biens, cette baisse est un peu plus faible. Par contre, la baisse est plus importante pour le studio. Le paramètre $\hat{\rho}$ baisse de 0,5936 à 0,3736 pour les transactions des studios en 2007 et de 0,4788 à 0,2273 en 1998. Le studio est un type d'appartement particulier et ce type de bien se trouve majoritairement à Paris. En effet, le studio est le bien est le plus recherché par les étudiants et les jeunes travailleurs. Cette particularité du studio est confirmée par l'analyse du degré de corrélation spatiale selon le nombre de pièces. La baisse du degré de corrélation est plus forte pour les biens d'une seule pièce, qui sont en majorité les studios. En 2007, la baisse du degré de corrélation est de 0,5932 à 0,3711 pour les biens d'une seule pièce mais seulement de 0,6203 à 0,5632 pour les biens de deux ou trois pièces et de 0,6152 à 0,4578 pour les biens de plus de trois pièces. Le même phénomène est observé si la condition de voisinage est la contiguïté d'ordre 4. Si la condition de voisinage est la distance seuil de 150 km, alors pour tous les types de biens, les degrés de corrélation

estimés à partir des données en dehors de l'Ile-de-France sont proches de zéro et même non significatif pour l'année 1998.

Les résultats obtenus de l'étude de corrélation spatiale des prix immobiliers en France montrent qu'il existe une autocorrélation spatiale entre les valeurs immobilières des différentes communes. De plus, la valeur d'un bien immobilier est corrélée non seulement à la valeur de son voisin le plus proche, mais aussi à la valeur des biens aux alentours. Le niveau de corrélation spatiale baisse considérablement si les données des transactions immobilières parisiennes sont exclues de la base de données. Cette baisse est plus importante si les données de toutes les communes en Ile-de-France sont ignorées. Dans certains cas, les prix immobiliers deviennent même non corrélés si les données des biens d'Ile-de-France sont exclues de la base de données. Ces résultats montrent que les prix des biens immobiliers localisés en province sont faiblement dépendants et que ces prix dépendent plutôt des prix immobiliers parisiens. Ces résultats permettent à confirmer le rôle dominant de Paris et de l'Ile-de-France dans la détermination des prix immobiliers en France. Paris est donc une ville dominante sur le marché immobilier français.

5.3. Existe-t-il un quartier dominant dans chaque ville ?

Nous poursuivons notre analyse à un niveau plus détaillé. Les données des transactions immobilières sont utilisées pour estimer le degré de corrélation spatiale dans chaque ville. La question posée est : *dans le cas d'une ville mono-centrique, le centre ville est-il un quartier dominant du marché immobilier de cette ville ? Dans une ville est polycentrique, comme Paris, Lyon, Marseille, quel est le quartier dominant et pourquoi ?*

Nous utilisons le même principe que dans la section précédente mais avec des unités d'étude plus petites. Les Tableau V.4, Tableau V.5 et Tableau V.6 montrent les degrés de corrélation spatiale pour Paris, Lyon et Marseille en 2007. Ces tableaux donnent les estimations de $\hat{\rho}$ obtenues en excluant tour à tour les données de chaque arrondissement. La variation du degré de corrélation permet de déterminer l'arrondissement qui joue un rôle dominant dans le marché immobilier de chaque ville. Si le degré de corrélation spatiale $\hat{\rho}$ baisse significativement après exclusion des données

d'un arrondissement, cela montre que cet arrondissement est dominant sur le marché immobilier de la ville. Inversement, si le degré de corrélation spatiale $\hat{\rho}$ augmente significativement après exclusion des données d'un arrondissement, cela signifie que les prix immobiliers de cet arrondissement ne sont pas corrélés aux prix immobiliers des autres arrondissements de la ville. Le marché immobilier de cet arrondissement pourrait alors avoir les caractéristiques spécifiques très différentes de celles des autres arrondissements.

5.3.1. Quartiers dominants à Paris

Le Tableau V.4 montre les résultats de l'estimation du degré de corrélation spatiale pour Paris. Le degré de corrélation $\hat{\rho}$ estimé à partir de l'ensemble des données parisiennes est 0,3244. Les paramètres $\hat{\rho}$ obtenus en enlevant les données de chaque arrondissement varient autour de 0,3 sauf pour certains arrondissements. Le degré de corrélation baisse à 0,2678, obtenu exclusion faite des données du 7ème arrondissement, et baisse à 0,2116 exclusion faite des données du 15ème arrondissement. Par contre, le degré de corrélation spatial augmente à 0,4950 exclusion faite des données du 17ème arrondissement et augmente à 0,3843 exclusion faite des données du 19ème arrondissement.

Pour simplifier la comparaison entre les différents coefficients de corrélation, nous calculons le rapport entre le degré de corrélation obtenu en enlevant les données de chaque arrondissement (ρ') et le degré de corrélation obtenu à partir de l'ensemble des transactions parisiennes (ρ_{Paris}). Si ce rapport, ρ'/ρ_{Paris}, est proche de 100% pour un arrondissement, cela signifie que le degré de corrélation obtenu en ignorant les données de cet arrondissement n'est pas différent de celui de l'ensemble de Paris. Une valeur faible de ce rapport signifie que le degré de corrélation spatiale baisse considérablement si les données de cet arrondissement sont exclues de la base de données. Ce qui indiquerait que les prix immobiliers des autres arrondissements sont fortement corrélés aux prix immobiliers de cet arrondissement, et donc que cet arrondissement présenterait un rôle dominant dans le marché immobilier résidentiel parisien. Inversement, une valeur de ρ'/ρ_{Paris} supérieure à 100% indique qu'un tel arrondissement présenterait des

caractéristiques spécifiques et que les prix immobiliers de cet arrondissement ne seraient pas corrélés à ceux des autres arrondissements.

Selon le Tableau V.4, ρ'/ρ_{Paris} est égal respectivement à 82,54% et 65,22% pour le 7ème arrondissement et le 15ème arrondissement. Cela nous conduit à établir le 15ème arrondissement comme le quartier directif du marché résidentiel parisien, le 7ème arrondissement présente aussi un rôle dominant mais moins important que celui du 15ème arrondissement. Notons que le 15ème arrondissement est bien connu des professionnels de l'immobilier comme une zone résidentielle, ce qui se confirme par la taille importante de ce marché résidentiel avec plus de 230 000 habitants[16]. Le 7ème arrondissement est l'un des arrondissements les plus chers et des plus recherchés par les investisseurs français et étrangers. Cet arrondissement est le marché résidentiel haut de gamme à Paris et présente un fort attrait notamment auprès d'une clientèle internationale fortunée[17]. Le niveau élevé de la demande et la taille importante du marché résidentiel ainsi que les informations données par les professionnels de l'immobilier nous permettent de confirmer que le 15ème arrondissement est un quartier dominant du marché résidentiel parisien de moyenne gamme et le 7ème arrondissement a rôle dominant dans le marché résidentiel parisien de haut de gamme.

Remarquons que le marché dominant n'est pas nécessairement l'arrondissement avec le prix moyen le plus élevé. Selon notre analyse, le 15ème arrondissement qui est considéré comme l'un des marchés dominants n'est pas l'arrondissement avec le prix moyen le plus élevé; par contre cet arrondissement est connu comme le quartier dominant du marché résidentiel parisien selon les spécialistes de l'immobilier. Ce rapport de corrélation est une méthode complémentaire permettant de mieux expliquer la situation actuelle du marché.

[16] Emmanuel Ducasse, « Conjecture et perspective immobilier », Direction études et observatoire immobilier, Crédit Foncier, mars 2012 [http://www.agent-immobilier-france.com/iledefrance/paris/index.htm].

[17] David TRAN, « Le marché résidentiel Haut de Gamme à Paris », CB Richard Ellis France View Point, octobre 2011 [http://www.cbre.fr/fr_fr/etudes/viewpoint].

Tableau V.4 : Degrés de corrélation estimés ($\hat{\rho}$) de Paris pour l'année 2007

Base de données	Nb Obs	Prix moyen (€/m2)	ρ' (Prix2007)		ρ'/ρ Paris
Paris	**28828**	**5335.82**	**32.44%**	*******	
Paris sans 1er ar	28559	6638.16	30.81%	***	94.98%
Paris sans 2ème ar	28403	5643.17	32.29%	***	99.53%
Paris sans 3ème ar	28176	6173.87	30.99%	***	95.53%
Paris sans 4ème ar	28386	6883.38	31.60%	***	97.41%
Paris sans 5ème ar	28135	6796.24	30.01%	***	92.51%
Paris sans 6ème ar	28199	7837.23	28.05%	***	86.46%
Paris sans 7ème ar	28091	7535.03	26.78%	***	82.55%
Paris sans 8ème ar	28100	6766.98	30.20%	***	93.10%
Paris sans 9ème ar	27676	5277.73	32.28%	***	99.50%
Paris sans 10ème ar	27327	4621.43	31.55%	***	97.25%
Paris sans 11ème ar	26543	4967.88	30.14%	***	92.91%
Paris sans 12ème ar	27245	4970.98	32.23%	***	99.35%
Paris sans 13ème ar	27214	5139.23	27.77%	***	85.61%
Paris sans 14ème ar	27371	5375.69	32.24%	***	99.39%
Paris sans 15ème ar	26052	5555.02	21.16%	***	65.23%
Paris sans 16ème ar	26397	6073.57	27.29%	***	84.13%
Paris sans 17ème ar	26151	5210.80	49.50%	***	152.59%
Paris sans 18ème ar	25779	4518.68	28.79%	***	88.75%
Paris sans 19ème ar	27059	4261.68	38.43%	***	118.46%
Paris sans 20ème ar	26869	4407.49	29.86%	***	92.04%

*Les symboles *, ** et *** désignent la significativité à 5%, 1% et 0,01%, respectivement.*

Base de données : Paris

Condition de voisinage : Distance de 250 mètres

Condition de pondération : Inverse de distance

Le rapport ρ'/ρ_{Paris} pour le 17ème arrondissement est de 152,57% et celui du 19ème arrondissement est 118,44%. L'exclusion de 17ème arrondissement des données étudiées conduit à une hausse considérable de l'autocorrélation spatiale. Le même résultat est obtenu si les données du 19ème arrondissement sont exclues de la base de l'étude. L'augmentation de $\hat{\rho}$ signifie que les 17ème et 19ème arrondissements contiennent des caractéristiques spécifiques à eux. Les biens résidentiels situés dans ces deux arrondissements se comportent différemment par rapport à ceux des autres arrondissements. Plusieurs éléments peuvent être évoqués comme caractéristique spécifique d'un tel arrondissement. Soit cet arrondissement présente un caractère typique

(par exemple, un quartier très dangereux ou un quartier d'immigrants), soit il contient des biens de type particulier (par exemple, des immeubles très anciens, des logements sociaux ou des villas). D'après le Tableau V.4, le prix moyen au mètre carré dans le 19ème arrondissement est le plus bas. Le rapport entre le nombre de logements sociaux et le nombre de résidences principales dans le 19ème arrondissement est 35,9%[18]. Ce volume élevé des logements sociaux peut expliquer son caractère spécifique car la valeur des logements sociaux est généralement déconnectée de la réalité du marché. Quant au 17ème arrondissement, son rôle spécifique peut s'expliquer par son hétérogénéité, cet arrondissement étant la combinaison de quartiers bourgeois de l'ouest parisien et des quartiers populaires du nord. Les appartements dans des immeubles de style haussmannien se trouvent dans la partie sud de l'arrondissement et les appartements plus petits localisés dans des rues moins larges se trouvent dans la partie nord de l'arrondissement[19]. Cette diversité de biens observée dans cet arrondissement peut être perçue comme sa caractéristique spécifique.

5.3.2. Quartiers dominants à Lyon

Le Tableau V.5 montre le degré de corrélation estimée des différents arrondissements de Lyon. Le degré de corrélation des prix immobiliers à l'intérieur de la commune de Lyon (appelé centre de Lyon) est égal à 0,2799. Le degré de corrélation des prix immobiliers est 0,1086 pour les communes autres que le centre de Lyon et 0,1954 pour tous les biens de l'agglomération lyonnaise. Ces trois résultats nous permettent de conclure qu'il y une corrélation des prix immobilier à l'intérieur du centre de Lyon. Par contre, les prix des biens situés en dehors du centre sont faiblement corrélés mais sont plutôt corrélés avec les prix des biens situés à l'intérieur du centre de Lyon. L'autocorrélation baisse considérablement si les données du centre de Lyon sont exclues de la base de données. Le centre de Lyon a donc un rôle directif sur les prix immobiliers des autres communes de l'agglomération lyonnaise.

[18] Francis Rol-Tanguy et al., « Les chiffres du logement social à Paris Début 2011 », Atelier Parisien d'Urbanisme, juillet 2011, [http://www.apur.org/sites/default/files/documents/logement-social-paris-2011.pdf].
[19] [http://www.seloger.com/immobilier/tout/immo-paris-17eme-75/] (page consultée le 6 septembre 2012).

En regardant le degré de corrélation estimé des différents arrondissements de Lyon, la plupart des rapports $\rho'/\rho_{Lyon\ Centre}$ sont proches de 100% sauf celui du 2ème arrondissement qui égal à 87,59%, celui du 5ème arrondissement qui est 114,91%, celui du 6ème arrondissement qui est 81,94% et celui du 9ème arrondissement qui est 113,22%. Ces résultats nous permettent d'établir le 2ème arrondissement et le 6ème arrondissement de Lyon comme les quartiers dominants du marché immobilier lyonnais. De plus, le rôle directif du 6ème arrondissement parait plus important que celui du 2ème arrondissement étant donné la baisse de corrélation plus importante pour le 6ème arrondissement. Les observations des professionnels du marché immobilier lyonnais corroborent bien nos résultats. Ces deux arrondissements sont en effet les arrondissements réputés bourgeois[20]. Le 2ème arrondissement est le quartier au coeur de la ville de Lyon avec la densité de commerces la plus forte et le quartier est considéré comme l'un des plus riches de la ville, où réside la bourgeoisie lyonnaise[21]. Le 6ème arrondissement de Lyon est un quartier résidentiel avec de très belles propriétés et appartements, le revenu par habitant le plus élevé[22]. Selon le Magazine *l'Express*, « le 6ème est historiquement le lieu où il faut habiter. C'est l'adresse prestigieuse de Lyon. Si l'on est riche, on réside dans cet arrondissement[23] ».

Le rapport $\rho'/\rho_{Lyon\ Centre}$ observé pour le 5ème arrondissement (114,91%) et le 9ème arrondissement (113,22%), considérablement supérieur à 100%, montre que ces deux arrondissements présentent des caractéristiques très spécifiques non cohérentes comparées aux autres quartiers de marché. En effet, le 5ème arrondissement est le centre historique de Lyon. Il comprend la colline de Fourvière et le quartier historique du Vieux Lyon[24]. C'est ce qui pourrait expliquer nos résultats qui montrent que le 5ème arrondissement présente un rôle spécifique parmi les différents arrondissements de Lyon.

[20] Feltin Michel, « Où vit-on le mieux à Lyon ? Neuf arrondissements au banc d'essai », L'Express, 26 avril 2001.
[21] Hêdre Guy, « Où vit-on le mieux à Lyon ? Classement général : 1ER , IIe arrondissement: Un hypercentre attirant », L'Express, 26 avril 2001.
[22] [http://www.salairemoyen.com/salaire-69386-Lyon_6e_Arrondissement.html] (page consultée le 6 septembre 2012).
[23] Francillon Claude, « Où vit-on le mieux à Lyon ? Classement général : 3E, VIe arrondissement : Le charme discret de la bourgeoisie », L'Express, 26 avril 2001.
[24] [http://www.seloger.com/immobilier/tout/immo-lyon-5eme-69/] (page consultée le 6 septembre 2012).

Le 9ème arrondissement quant à lui est le quartier le plus pauvre[25]. De plus, puisque cet arrondissement est le plus éloigné du centre ville, le moins desservi par les transports en commun cela peut être une contrainte qui fait que les prix dans cet arrondissement ne soient corrélés à ceux des autres arrondissements.

Tableau V.5 : Degrés de corrélation estimés ($\hat{\rho}$) de Lyon pour l'année 2007

Base de données	Nb Obs	Prix moyen (€/m2)	ρ' (Prix2007)		ρ'/ρ centre de Lyon
Agglomération Lyon	10312	2272.76	19.54%	***	
Centre de Lyon	**4658**	**2462.45**	**27.99%**	***	
Agglomération sans Lyon centre	5654		10.86%	***	
Lyon sans 1er ar	4321	2497.49	27.69%	***	98.94%
Lyon sans 2ème ar	4399	2659.54	24.51%	***	87.59%
Lyon sans 3ème ar	3712	2426.08	27.23%	***	97.28%
Lyon sans 4ème ar	4266	2665.40	26.57%	***	94.94%
Lyon sans 5ème ar	4224	2291.74	32.16%	***	114.91%
Lyon sans 6ème ar	4173	2732.34	22.93%	***	81.94%
Lyon sans 7ème ar	4028	2279.81	26.30%	***	93.96%
Lyon sans 8ème ar	4029	2343.06	28.12%	***	100.46%
Lyon sans 9ème ar	4112	2508.90	31.69%	***	113.22%

*Les symboles *, ** et *** désignent la significativité à 5%, 1% et 0,01%, respectivement.*

Base de données : Lyon

Condition de voisinage : Distance de 530 mètres

Condition de pondération : Inverse de distance

5.3.3. Quartiers dominants à Marseille

Marseille est une autre ville de France qui est subdivisée en plusieurs arrondissements. Il existe au total 16 arrondissements à Marseille. Le Tableau V.6 détaille le degré de corrélation obtenu à partir des données de Marseille. Les arrondissements pour lesquels le rapport de corrélation est sensiblement différent de 100% sont les suivants : le 7ème arrondissement avec 84,67%, le 8ème arrondissement avec 89,84%, le

[25] Feltin Michel, op.cit. (note 20).

9ème arrondissement avec 105,29%, le 13ème arrondissement avec 105,10% et le 16ème arrondissement avec 105,11%.

D'après ces résultats, le 7ème arrondissement et le 8ème arrondissement sont les deux arrondissements dominants de Marseille. Les prix immobiliers de Marseille sont essentiellement corrélés aux prix des biens immobiliers localisés dans ce quartier. Pour confirmer ces résultats, on peut analyser de plus près les caractéristiques du marché immobilier marseillais. L'étude du marché marseillais montre que les quartiers résidentiels les plus aisés et les plus recherchés pour la bonne qualité de vie sont les 7ème et 8ème arrondissements[26].

Le rapport élevé que l'on observe pour les 9ème arrondissement, 13ème arrondissement et 16ème arrondissement montre que ces trois arrondissements présentent des caractéristiques très spécifiques. Selon les professionnels de l'immobilier marseillais[27], certains quartiers du 9ème arrondissement offrent beaucoup plus de maisons ou de villas que d'appartements ce qui peut être considéré comme une caractéristique spécifique de ces quartiers. Les prix immobiliers dans le 13ème arrondissement ne sont pas corrélés aux autres quartiers de la ville car dans cet arrondissement de quartiers résidentiels et de cités HLM sensibles[28], la présence de logements sociaux pourrait expliquer la spécificité de cet arrondissment. Le 16ème est l'arrondissement dont la spécificité est qu'il est le plus éloigné du centre ville.

[26] Pierre Falga, « Où vit-on le mieux à Marseille 16 arrondissements passés au crible », *L'Express*, 25 octobre 2001.

[27] [http://www.seloger.com/immobilier/tout/immo-marseille-9eme-13/] (page consultée le 6 septembre 2012).

[28] Leras Marc « Où vit-on le mieux à Marseille ? XIIIème arrondissement : Une perle nommée Château-Gombert », L'Express, 25 octobre 2001.

Tableau V.6 : Degrés de corrélation estimés ($\hat{\rho}$) de Marseille pour l'année 2007

Base de données	Nb Obs	Prix moyen (€/m2)	ρ' (Prix2007)		ρ'/ρ centre de Marseille
Agglomération Marseille	5849	2440.88	24.47%	***	
Centre de Marseille	**3969**	**2321.01**	**23.68%**	*******	
Marseille sans le centre	1378		26.39%	**	
Marseille sans 1er ar	3742	1969.88	22.45%	***	94.81%
Marseille sans 2ème ar	3845	2251.98	22.93%	***	96.83%
Marseille sans 3ème ar	3763	1881.74	22.05%	***	93.11%
Marseille sans 4ème ar	3649	2112.27	23.41%	***	98.87%
Marseille sans 5ème ar	3639	2253.38	23.70%	***	100.11%
Marseille sans 6ème ar	3686	2357.51	23.38%	***	98.74%
Marseille sans 7ème ar	3716	2837.68	20.05%	***	84.67%
Marseille sans 8ème ar	3449	2969.44	21.27%	***	89.84%
Marseille sans 9ème ar	3620	2476.34	24.93%	***	105.29%
Marseille sans 10ème ar	3729	2276.84	23.60%	***	99.67%
Marseille sans 11ème ar	3831	2290.00	24.74%	***	104.48%
Marseille sans 12ème ar	3688	2426.46	23.67%	***	99.96%
Marseille sans 13ème ar	3800	2076.72	24.88%	***	105.10%
Marseille sans 14ème ar	3743	1851.98	23.07%	***	97.44%
Marseille sans 15ème ar	3747	1716.29	22.86%	***	96.56%
Marseille sans 16ème ar	3888	2348.13	24.89%	***	105.11%

*Les symboles *, ** et *** désignent la significativité à 5%, 1% et 0,01%, respectivement.*

Base de données : Marseille

Condition de voisinage : Distance de 1500 mètres

Condition de pondération : Inverse de distance

L'analyse des données de Lyon et Marseille confirment qu'il existe un ou plusieurs arrondissements dominants dans le marché immobilier de chaque ville. Ces résultats trouvés correspondent à la réalité du marché immobilier de chaque ville. Notons qu'il n'existe pas à ce jour d'indicateur macroéconomique permettant d'identifier le quartier directeur de chaque ville. En comparant le rapport de corrélation au niveau de prix moyen de chaque arrondissement, nous montrons que le niveau du prix immobilier seul ne permet pas de définir le rôle dominant d'un arrondissement donné. Le rapport de corrélation utilisé dans notre analyse peut être perçu comme un indicateur complémentaire permettant de mieux expliquer la réalité du marché immobilier.

Tableau V.7 : Degrés de corrélation estimés ($\hat{\rho}$) des neuf agglomérations (avec et sans les données de centre ville) de France pour l'année 1998 et 2007

Base de données	Nb Obs	Prix moyen (€/m2)	ρ (Prix1998)		Base de données	Nb Obs	Prix moyen (€/m2)	ρ (Prix2007)	
Agglomération Bordeaux	2929	1063.75	24.96%	***	Agglomération Bordeaux	3898	2490.99	24.56%	***
Bordeaux sans le centre	1513	1091.49	20.43%	***	Bordeaux sans le centre	2038	2411.12	20.35%	***
Centre de Bordeaux	1416	1034.12	12.81%	***	Centre de Bordeaux	1860	2578.51	19.93%	***
Agglomération Lille	2127	1200.20	27.44%	***	Agglomération Lille	2941	2470.21	40.44%	***
Lille sans le centre	854	1150.66	13.38%	***	Lille sans le centre	1162	2339.82	34.74%	***
Centre de Lille	1273	1233.43	31.13%	***	Centre de Lille	1779	2555.37	37.36%	***
Montpellier Agglomération	2330	1223.08	33.29%	***	Agglomération Montpellier	4094	2835.75	35.56%	***
Montpellier sans le centre	420	1243.41	42.33%	***	Montpellier sans le centre	1095	3204.43	40.27%	***
Montpellier centre	1916	1218.63	29.66%	***	Centre de Montpellier	2174	2705.51	27.97%	***
Agglomération Nantes	3262	1119.63	26.36%	***	Agglomération Nantes	4094	2561.89	29.31%	***
Nantes sans le centre	867	1046.03	17.33%	***	Nantes sans le centre	1095	2358.67	23.60%	***
Centre de Nantes	2395	1146.27	23.87%	***	Centre de Nantes	2999	2636.09	26.41%	***
Agglomération Nice	10379	1517.12	35.76%	***	Agglomération Nice	13294	3987.34	39.37%	***
Nice sans le centre	5824	1640.68	44.24%	***	Nice sans le centre	7917	4271.90	48.15%	***
Centre de Nice	4555	1359.13	30.93%	***	Centre de Nice	5377	3568.36	38.37%	***
Agglomération Orléans	1653	1202.71	34.54%	***	Agglomération Orléans	1683	2055.10	23.88%	***
Orléans sans le centre	616	1139.20	45.95%	***	Orléans sans le centre	590	2003.51	30.27%	***
Centre d'Orléans	1037	1240.44	17.66%	***	Centre d'Orléans	1093	2082.95	17.72%	***
Agglomération Rennes	1907	1188.05	26.24%	***	Agglomération Rennes	3667	2449.04	28.33%	***
Rennes sans le centre	396	1271.61	12.36%	***	Rennes sans le centre	1538	2401.49	18.33%	***
Centre de Rennes	1511	1166.15	27.41%	***	Centre de Rennes	2129	2483.40	29.89%	***
Agglomération Strasbourg	2461	1325.62	32.01%	***	Agglomération Strasbourg	4132	2380.08	34.26%	***
Strasbourg sans le centre	1104	1361.38	14.47%	***	Strasbourg sans le centre	1954	2358.03	18.98%	***
Centre de Strasbourg	1357	1296.52	42.65%	***	Centre de Strasbourg	2178	2399.86	40.24%	***
Agglomération Toulouse	3651	1162.76	36.91%	***	Agglomération Toulouse	5649	2767.40	27.62%	***
Toulouse sans le centre	903	1310.03	22.63%	***	Toulouse sans le centre	2188	2789.40	9.62%	***
Centre de Toulouse	2748	1114.37	39.66%	***	Centre de Toulouse	3461	2753.50	40.48%	***

*Les symboles *, ** et *** désignent la significativité à 5%, 1% et 0,01%, respectivement.*

5.3.4. Quartiers dominants des villes en France

Pour les 9 villes restantes qui ne sont pas subdivisées en arrondissements, l'autocorrélation spatiale est estimée avec et sans les données du centre ville de chaque agglomération. Cela permet de vérifier si le centre ville a un rôle dominant dans le marché immobilier de chacune de ces agglomérations. Parmi ces 9 agglomérations de France, nos résultats montrent que le centre ville est une zone dominante du marché immobilier de ces agglomérations sauf pour Montpellier, Nice et Orléans. L'explication de ces résultats nécessite une meilleure connaissance de l'environnement sociologique de ces villes.

6. Conclusion

Avant de faire leurs choix d'investissement ou d'achat, les investisseurs en immobilier ou les acquéreurs potentiels se posent la question de savoir quel est le quartier dominant du marché immobilier de la ville. Cette question est utile pour observer l'évolution de marché et estimer la valeur du bien. Si la ville en question est une ville mono-centrique, le centre ville est sans doute la région dominante du marché immobilier mais si la ville en question correspond à une ville polycentrique comme Paris, Lyon ou Marseille, il y'a lieu de se poser la question de savoir comment définir le quartier dominant. Les professionnels de l'immobilier peuvent désigner le quartier dominant en se basant sur leurs connaissances du terrain ou sur le développement historique de la ville. Il reste cependant à savoir comment on peut confirmer, statistiquement, ce rôle dominant d'un quartier donné.

Ce travail définit le quartier dominant en fonction du degré de corrélation spatiale des prix immobiliers. Si les données de tel quartier sont enlevées de la base de données, le degré de corrélation spatiale baisse sensiblement, alors ce quartier présente un rôle directif dans l'effet de diffusion des prix immobiliers. Ce quartier est donc considéré comme un quartier dominant. Inversement, si le degré de corrélation spatiale augmente sensiblement lorsque les données d'un quartier donné sont exclues de la base de données,

alors ce quartier présente une caractéristique spécifique qui est différente par rapport aux autres quartiers de la ville.

Le centre ville apparait comme la zone dominante du marché immobilier de la ville pour les neuf agglomérations de France (Bordeaux, Lille, Montpellier, Nantes, Nice, Orléans, Rennes, Strasbourg, Toulouse), pour toutes les villes sauf Montpellier, Nice et Orléans. Pour les trois grandes villes de France, à savoir Paris, Lyon et Marseille, nous mettons en exergue l'arrondissement dominant de chaque ville. Pour Paris, nos résultats montrent que les $7^{ème}$ et $15^{ème}$ arrondissements, qui sont bien connus par les professionnels immobiliers comme des zones résidentielles, sont les arrondissements dominants, et les $17^{ème}$ et $19^{ème}$ arrondissements présentent des caractéristiques spécifiques par rapport aux autres arrondissements. La diversité des biens dans le $17^{ème}$ et le pourcentage important de logements sociaux dans le $19^{ème}$ peuvent expliquer leurs spécificités. Pour Lyon, les $2^{ème}$ et $6^{ème}$ arrondissements qui sont les quartiers résidentiels bourgeois présentent un rôle dominant sur marché résidentiel lyonnais, et l'arrondissement du Vieux-Lyon (le $5^{ème}$ arrondissement) ainsi que l'arrondissement le plus éloigné du centre (le $9^{ème}$ arrondissement) se distinguent par ces caractéristiques spécifiques. Les résultats de Marseille montrent que les $7^{ème}$ et $8^{ème}$ arrondissements sont les arrondissements dominants du marché, et les $9^{ème}$, $13^{ème}$ et $16^{ème}$ arrondissements sont les arrondissements avec des caractéristiques spécifiques.

Nos résultats confirment bien les observations que nous pouvons faire sur ces différents marchés immobiliers. Ce travail permet ainsi de confirmer, statistiquement et économétriquement, le caractère dominant de certains tel qu'habituellement cités par les experts immobiliers. Il est difficile d'identifier un indicateur macroéconomique existant qui expliquerait ce rôle dominant. Selon nos résultats, le prix moyen au mètre carré ne peut pas caractériser le quartier dominant, et ce n'est le cas ni pour le revenu des habitants, ni pour le taux de criminalité. Nous montrons que la différence de degrés de corrélation peut en revanche être considérée comme un indicateur intéressant pour déterminer le quartier dominant.

Afin de compléter ce travail, le test de comparaison entre deux degrés de corrélations est envisageable, mais il n'existe à notre connaissance à ce jour aucun test

spécifique du degré de corrélation spatiale. La transformation *z* de Fisher qui permet de comparer deux paramètres de corrélation est valable uniquement pour les corrélations de Pearson.

CONCLUSION GENERALE

Cette thèse s'intéresse à l'évaluation des biens immobiliers en présence d'une dépendance spatiale des prix. Les deux approches de la statistique spatiale que sont la géostatistique et l'économétrie spatiale sont utilisées pour analyser cette dépendance spatiale dans plusieurs contextes et à plusieurs fins : afin de donner une bonne estimation de prix, de fournir une meilleure prévision de la valeur immobilière, de construire des indices, de déterminer des segmentations de marché et de considérer l'influence d'une nouvelle infrastructure sur le prix immobilier.

Après la présentation des éléments de ces deux approches, des hypothèses posées lors de la construction du modèle, des méthodes d'estimation, différents points sont analysés : leurs différences, leurs ressemblances, les avantages et les inconvénients de chaque approche ainsi que les applications de ces approches dans le cas d'une étude immobilière.

Géostatistique – Économétrie spatiale : les ressemblances et les différences ?

Dans le cas de présence d'une dépendance spatiale, deux perspectives permettent de formaliser ce phénomène : soit le modèle hédonique extensif, en ajoutant les caractéristiques spatiales en tant que variables explicatives au modèle de régression, soit des méthodes statistiques cherchant à définir directement le degré de dépendance entre les observations. La géostatistique et l'économétrie spatiale appartiennent à cette deuxième perspective. Ces deux approches déterminent le degré de dépendance entre les observations par l'intermédiaire de la matrice de variance-covariance. Cette matrice est utilisée, par la suite, dans la régression par les moindres carrés généralisés afin d'obtenir des estimateurs non biaisés.

Par contre, ces deux approches se différencient par la méthode appliquée pour déterminer le degré de dépendance ainsi que par les hypothèses émises pour développer ces techniques. Sous l'hypothèse de la continuité et de la stationnarité de la distribution spatiale, la géostatistique estime la covariance en fonction de la distance entre les observations, on obtient ainsi le covariogramme. Celui-ci est utilisé ensuite pour déterminer la matrice de variance-covariance utilisée dans la régression des moindres carrés généralisés. L'économétrie spatiale analyse la dépendance spatiale d'une autre

façon. Elle travaille directement avec la matrice d'interaction entre les observations ; on détermine le voisinage de chaque observation et on accorde des poids différents selon le voisinage. La matrice de poids est intégrée à l'équation de régression afin d'estimer les coefficients non biaisés du modèle de prix hédoniste.

Géostatistique – Économétrie spatiale : les avantages et les inconvénients d'une approche par rapport à l'autre ?

La géostatistique et l'économétrie spatiale sont donc toutes les deux des approches de statistique spatiale permettent d'analyser la dépendance spatiale des prix immobiliers mais chaque méthode présente des avantages et des inconvénients.

Les deux hypothèses de la géostatistique paraissent assez contraignantes. L'hypothèse de la continuité spatiale signifie que les indicateurs de localisation d'observations se distribuent de façon aléatoire dans l'espace continu. Plus précisément, la base de données ne devrait pas être coupée par une frontière ou une barrière. Sous cette réserve, cette hypothèse semble généralement valable si le nombre des données collectées est important. Par contre, dans le cas des données immobilières, il est tout à fait possible que la ville soit découpée en plusieurs segments par des frontières naturelles ou politiques. Dans ce cas, l'analyse géostatistique doit s'appliquer à chaque segment. La deuxième hypothèse contraignante de la géostatistique est la stationnarité spatiale. Selon cette hypothèse, le variogramme d'un processus stationnaire doit avoir la même allure pour n'importe quelle segmentation des données. Or le caractère très hétérogène des biens immobiliers a tendance à s'opposer à cette hypothèse, peu souvent vérifiée en pratique. Notre analyse géostatistique des données de transactions résidentielles parisiennes montre en effet qu'il est difficile de considérer que les variogrammes des différents segments soient stationnaires. Le variogramme et la portée estimés varient selon la région d'étude et les années d'étude. Dans le cas de l'analyse géostatistique des données immobilières, il n'est pas raisonnable de faire cette hypothèse de la stationnarité d'une manière automatique. Un autre inconvénient de cette méthode pour l'immobilier est qu'elle nécessite des informations précises sur la localisation de chaque observation, comme les coordonnées géographiques ou les coordonnées cartésiennes afin de pouvoir calculer la

distance entre les observations. Or, si les bases immobilières sont en cours de géolocalisation, cela n'est pas encore une règle systématique.

La géostatistique définit souvent la covariance en fonction uniquement de la distance entre les observations (hypothèse d'isotropie). Cette approche est avantageuse par rapport à l'économétrie spatiale car cette simplification permet de faciliter l'estimation et de réduire le temps de calcul. Toutefois, cette hypothèse simplificatrice ne peut pas être faite non plus directement. Il est en effet envisageable que la covariance entre les observations dépende aussi de l'orientation entre celles-ci.

Enfin, l'approche géostatistique permettant de déterminer la distance à partir de laquelle la covariance devient standard, un autre avantage de cette optique est que cette distance limite peut servir pour la segmentation de marché.

L'économétrie spatiale d'attache de sont coté à déterminer la matrice d'interaction spatiale des observations, qui sera intégrée au modèle de régression hédonique. L'approche de l'économétrie spatiale présente un avantage parce qu'elle permet de déterminer cette interaction de plusieurs façons. Dans le cas de la géostatistique, la structure de corrélation entre les observations est simplifiée en posant une relation inverse à la distance quelque soit la distribution spatiale des données ; par contre, pour l'approche d'économétrie spatiale, l'interaction est définie soit par la contiguïté, soit par une condition de distance. De plus, la prise en compte de l'autocorrélation dans les modèles économétriques spatiaux peut s'effectuer de plusieurs manières : soit par des variables endogènes décalées, soit par des variables exogènes décalées, soit par une autocorrélation spatiale des erreurs. Cela donne un certain avantage par rapport à l'approche géostatistique qui analyse la dépendance spatiale uniquement à partir des résidus de l'estimation.

Néanmoins, cette approche dépend du choix ex-ante de la condition de voisinage et du choix des poids accordés aux observations voisines. De plus, si le nombre des observations (n) est important, la matrice de taille $n \times n$ est difficilement maniable et nécessite un équipement puissant pour l'élaborer.

Géostatistique – Économétrie spatiale : Quelle approche pour quel contexte immobilier ?

Il n'existe pas de réponse exacte et définitive à cette question, tout dépend des données étudiées, de leur distribution dans l'espace, de la source de l'autocorrélation spatiale analysée, de l'information de localisation, du logiciel disponible et de l'objectif de l'étude. Si les deux méthodes peuvent être utilisées pour étudier la dépendance spatiale des prix immobiliers, il existe cependant des points sur lesquels il convient d'être vigilent.

Avant d'utiliser une analyse géostatistique, une étape indispensable est de vérifier si les données sont distribuées dans un espace continu, si les variogrammes obtenus pour les différents quartiers sont bien stationnaires et si la corrélation spatiale entre les prix des biens immobiliers dépend uniquement de la distance entre leur localisation. Cette analyse géostatistique est normalement appliquée aux résidus de l'estimation hédonique. Certaines études précédentes font remarquer que les caractéristiques spatiales ne doivent pas être incluses dans le modèle de prix hédoniste afin de laisser les résidus capter pleinement la spatialisation de la structure; cependant l'étude des variogrammes non stationnaires montre qu'une fois ajouté l'indicateur du quartier à la régression hédonique, cela permet de réduire sensiblement la fluctuation des variogrammes estimés. Ces résultats nous ont conduits à diviser l'influence de la localisation sur le prix immobilier en deux niveaux. Le premier correspond aux « effets de quartier », les logements situés dans un même quartier partageraient des caractéristiques « de quartier ». Cet effet du quartier serait une influence qualifiée de *macro*, c'est-à-dire commune à tous les biens du quartier. Le deuxième niveau, le niveau *micro*, correspondrait aux « effets de contiguïté », c'est à dire, l'influence du prix d'un bien sur les prix des autres biens mitoyens ou très proches. L'indicateur du quartier ajouté à la régression hédonique permettrait ainsi de capter partiellement les effets macro tandis que les effets de contiguïté seraient adressés par le travail sur les résidus de la régression hédonique.

Concernant l'approche d'économétrie spatiale, elle requière donc un choix ex-ante de la matrice d'interaction et du modèle spatial. La matrice de voisinage peut être définie par la contiguïté ou la condition de distance. La contiguïté permet de limiter le nombre de voisinages mais elle nécessite la détermination de l'ordre de contiguïté. Si la base de données est très dense, la contiguïté d'ordre 1 n'est pas suffisante. La valeur d'un bien

immobilier est corrélée non seulement à la valeur du bien voisin le plus proche, mais aussi à la valeur des biens qui se trouvent autour. Si la base de données est diffuse, la contiguïté d'ordre 1 pourrait considérer comme voisins des couples d'observations qui seraient trop éloignées. La matrice de distance permet alors de limiter la distance maximale entre les voisinages, mais il existe un risque d'avoir une observation isolée, sans voisins, si la distance seuil n'est pas suffisamment élevée ou si les données sont très étalées. Parmi les différents modèles spatiaux disponibles, seul le modèle autorégressif spatial et le modèle d'erreurs spatial sont utilisés dans l'étude immobilière. Certains modèles tels que celui à variables exogènes décalées ou le modèle de Durbin spatial peuvent être utilisés dans le cadre des études de dépendance des prix immobiliers. Le choix entre les différents modèles dépend de la source de l'autocorrélation spatiale prise en compte dans l'étude. Le modèle des variables endogènes décalées est adapté si c'est le processus d'évaluation du bien immobilier qui est la cause de l'autocorrélation spatiale. Le modèle à variables exogènes décalées sera utilisé si la source considérée est la similarité des biens, le modèle d'erreurs spatial sera appliqué dans le cas où la dépendance est causée par la mauvaise spécification du modèle de l'estimation.

Cette thèse a mise en œuvre différents exemples d'analyse issue de la statistique spatiale, selon les différentes sources de la dépendance. La géostatistique a permis d'analyser notamment les résidus de l'estimation hédonique, cette approche a considérée que la dépendance spatiale se présentait uniquement par les résidus (Section 5 du Chapitre 3). En cas de mauvaise spécification du modèle hédonique certaines variables explicatives, qui comportent de la dépendance spatiale, peuvent ne pas être prises en compte dans le modèle. L'analyse géostatistique a pu ainsi être améliorée en ajoutant l'indicateur du quartier au modèle hédonique (Section 6 du Chapitre 3). La dernière partie de ce travail est basée sur l'idée que le processus d'évaluation immobilière crée l'effet de diffusion des prix immobiliers (Chapitre 5). Le degré de corrélation est alors utilisé pour déterminer le quartier dominant du marché immobilier de chaque ville. Ce quartier dominant serait utile pour l'estimation des prix immobiliers, pour la présentation de l'évolution indicielle des prix ou pour la détermination d'une référence de prix.

BIBLIOGRAPHIE

[1] Adair, Alastair, Stanley McGreal, Austin Smyth, James Cooper, et Tim Ryley, (2000), "House prices and accessibility: The testing of relationships within the belfast urban area", *Housing Studies* 15, 699-716.

[2] Anselin, L., et S.J. Rey, (2009). *Perspectives on spatial data analysis* (Springer).

[3] Anselin, Luc, (1988). *Spatial econometrics : Methods and models* (Kluwer Academic Publishers, Dordrecht ; Boston).

[4] Anselin, Luc, (1998), "Gis research infrastructure for spatial analysis of real estate markets", *Journal of Housing Research* 9.

[5] Anselin, Luc, R. J. G. M. Florax, et Sergio J. Rey, (2004). *Advances in spatial econometrics: Methodology, tools and applications* (Springer, Berlin).

[6] Armstrong, Margaret, et Jacques Carignan, (1997). *Géostatistique linéaire application au domaine minier* (les Presses de l'École des mines, Paris).

[7] Asli, Mustapha, et Denis Marcotte, (1995), "Comparaison de différentes approches de krigeage dans un contexte multivariable", *Mathematical Geology* 27.

[8] Aten, Bettina, (1996), "Evidence of spatial autocorrelation in international prices", *Review of Income and Wealth* 42, 149-163.

[9] Atkinson, P. M., et C. D. Lloyd, (2007), "Non-stationary variogram models for geostatistical sampling optimisation: An empirical investigation using elevation data", *Computers & Geosciences* 33, 1285-1300.

[10] Bailey, Trevor C., et Anthony C. Gatrell, (1995). *Interactive spatial data analysis* (Longman Scientific and Technical, Harlow).

[11] Barnes, Randal, (1991), "The variogram sill and the sample variance", *Mathematical Geology* 23, 673-678.

[12] Basu, Sabyasachi, et Thomas G. Thibodeau, (1998), "Analysis of spatial autocorrelation in house prices", *The Journal of Real Estate Finance and Economics* 17, 61-85.

[13] Bender, Bruce, et Hae-Shin Hwang, (1985), "Hedonic housing price indices and secondary employment centers", *Journal of Urban Economics* 17, 90-107.

[14] Besner, Claude, (2002), "A spatial autoregressive specification with a comparable sales weighting scheme", *Journal of Real Estate Research* 24.

[15] Bible, Douglas S., et Cheng-Ho Hsieh, (1996), "Applications of geographic information systems for the analysis of apartment rents", *Journal of Real Estate Research* 12.

[16] Bivand, Roger, (2012), " Spatial dependence: Weighting schemes, statistics and model", *Package Spdep in R*.

[17] Bourassa, Steven C., Foort Hamelink, Martin Hoesli, et Bryan D. MacGregor, (1999), "Defining housing submarkets", *Journal of Housing Economics* 8, 160-183.

[18] Bourassa, Steven, Eva Cantoni, et Martin Hoesli, (2007), "Spatial dependence, housing submarkets, and house price prediction", *The Journal of Real Estate Finance and Economics* 35, 143-160.

[19] Bowen, William M., Brian A. Mikelbank, et Dean M. Prestegaard, (2001), "Theoretical and empirical considerations regarding space in hedonic housing price model applications", *Growth and Change* 32, 466-490.

[20] Bowes, David R., et Keith R. Ihlanfeldt, (2001), "Identifying the impacts of rail transit stations on residential property values", *Journal of Urban Economics* 50, 1-25.

[21] Brasington, David, (1999), "Which measures of school quality does the housing market value?", *Journal of Real Estate Research* 18.

[22] Bruckel, Denis, François Cusin, Claire Juillard, et Arnaud Simon, (2009), "Note de conjoncture immobilière", in Notaires de France, ed.

[23] Caloz, Régis, et Claude Collet, (2011). *Analyse spatiale de l'information géographique* (Presses polytechniques et universitaires romandes [diff. Geodif], Lausanne [Paris]).

[24] Can, Ay, et Isaac Megbolugbe, (1997), "Spatial dependence and house price index construction", *The Journal of Real Estate Finance and Economics* 14, 203-222.

[25] Can, Ayse, (1990), "The measurement of neighborhood dynamics in urban house prices", *Economic Geography* 66, 254-272.

[26] Can, Ayse, (1992), "Specification and estimation of hedonic housing price models", *Regional Science and Urban Economics* 22, 453-474.

[27] Cano-Guervós, Rafael, Jorge Chica-Olmo, et José A. Hermoso-Gutiérrez, (2003), "A geo-statistical method to define districts within a city", *The Journal of Real Estate Finance and Economics* 27, 61-85.

[28] Case, Anne C., Harvey S. Rosen, et James R. Hines Jr, (1993), "Budget spillovers and fiscal policy interdependence: Evidence from the states", *Journal of Public Economics* 52, 285-307.

[29] Case, Bradford, John Clapp, Robin Dubin, et Mauricio Rodriguez, (2004), "Modeling spatial and temporal house price patterns: A comparison of four models", *The Journal of Real Estate Finance and Economics* 29, 167-191.

[30] Chauvet, P., (1999). *Aide-mémoire de géostatistique linéaire* (Presses de l'École des Mines).

[31] Cherry, Steve, (1997), "Non-parametric estimation of the sill in geostatistics", *Environmetrics* 8, 13-27.

[32] Clapp, John M., (2003), "A semiparametric method for valuing residential locations: Application to automated valuation", *The Journal of Real Estate Finance and Economics* 27, 303-320.

[33] Clapp, John M., (2004), "A semiparametric method for estimating local house price indices", *Real Estate Economics* 32, 127-160.

[34] Clapp, John M., et Yazhen Wang, (2006), "Defining neighborhood boundaries: Are census tracts obsolete?", *Journal of Urban Economics* 59, 259-284.

[35] Conway, Delores, Christina Li, Jennifer Wolch, Christopher Kahle, et Michael Jerrett, (2010), "A spatial autocorrelation approach for examining the effects of urban greenspace on residential property values", *The Journal of Real Estate Finance and Economics* 41, 150-169.

[36] Corstanje, R., S. Grunwald, et R. M. Lark, (2008), "Inferences from fluctuations in the local variogram about the assumption of stationarity in the variance", *Geoderma* 143, 123-132.

[37] Cressie, Noel, (1988), "Spatial prediction and ordinary kriging", *Mathematical Geology* 20, 405-421.

[38] Cressie, Noel A. C., (1991). *Statistics for spatial data* (J. Wiley, New York).

[39] Debrezion, Ghebreegziabiher, Eric Pels, et Piet Rietveld, (2007), "The impact of railway stations on residential and commercial property value: A meta-analysis", *The Journal of Real Estate Finance and Economics* 35, 161-180.

[40] Des Rosiers, F., M. Thériault, et P.Y. Villeneuve, (2000), "Sorting out access and neighbourhood factors in hedonic price modelling : An application to the quebec city metropolitan area", *The Journal of Property Investment and Finance* 18.

[41] Droesbeke, J.J., M. Lejeune, et G. Saporta, (2006a). *Analyse statistique des données spatiales* (Technip).

[42] Droesbeke, Jean-Jacques, Michel Lejeune, et Gilbert Saporta, (2006b). *Analyse statistique des données spatiales* (Éd. Technip, Paris).

[43] Drukker, D. M., I. R. Prucha, et R. Raciborski, (2011), "An ml and gs2sls estimator for the spatial autoregressivemodel", *Technical report, Stata*.

[44] Dubin, R. A., et C. H. Sung, (1990), "Specification of hedonic regressions - nonnested tests on measures of neighborhood quality", *Journal of Urban Economics* 27, 97-110.

[45] Dubin, Robin, (2003), "Robustness of spatial autocorrelation specifications: Some monte carlo evidence", *Journal of Regional Science* 43, 221-248.

[46] Dubin, Robin A., (1988), "Estimation of regression coefficients in the presence of spatially autocorrelated error terms", *The Review of Economics and Statistics* 70.

[47] Dubin, Robin A., (1992), "Spatial autocorrelation and neighborhood quality", *Regional Science and Urban Economics* 22, 433-452.

[48] Dubin, Robin A., (1998), "Predicting house prices using multiple listings data", *The Journal of Real Estate Finance and Economics* 17, 35-59.

[49] Dubin, Robin A., et Chein-Hsing Sung, (1987), "Spatial variation in the price of housing: Rent gradients in non-monocentric cities", *Urban Studies* 24, 193-204.

[50] Dubin, Robin, R. Kelley Pace, et Thomas G. Thibodeau, (1999), "Spatial autoregression techniques for real estate data", *Journal of Real Estate Literature* 7, 79-96.

[51] Dunse, Neil, et Colin Jones, (1998), "A hedonic price model of office rents", *Journal of Property Valuation and Investment* 16, 297-312.

[52] Ekström, Magnus, (2008), "Subsampling variance estimation for non-stationary spatial lattice data", *Scandinavian Journal of Statistics* 35, 38-63.

[53] Ekström, Magnus, et Sara Sjöstedt-De Luna, (2004), "Subsampling methods to estimate the variance of sample means based on nonstationary spatial data with varying expected values", *Journal of the American Statistical Association* 99, 82-95.

[54] Fernandez-Aviles, Gema, Roman Minguez, et Jose-Maria Montero, (2012), "Geostatistical air pollution indexes in spatial hedonic models: The case of madrid, spain", *Journal of Real Estate Research* 34.

[55] Fernández-Casal, Rubén, Wenceslao González-Manteiga, et Manuel Febrero-Bande, (2003), "Flexible spatio-temporal stationary variogram models", *Statistics and Computing* 13, 127-136.

[56] Fik, Timothy J., David C. Ling, et Gordon F. Mulligan, (2003), "Modeling spatial variation in housing prices: A variable interaction approach", *Real Estate Economics* 31, 623-646.

[57] Fingleton, Bernard, (2008), "A generalized method of moments estimator for a spatial model with moving average errors, with application to real estate prices", *Empirical Economics* 34, 35-57.

[58] Fisher, R. A., (1915), "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population", *Biometrika* 10, 507-521.

[59] Fisher, R. A., (1921), "On the probable error of a coefficient of correlation deduced from a small sample", *Metron* 1, 3-32.

[60] Gaetan, Carlo, Xavier Guyon, et Kevin Bleakley, (2010), "Spatial statistics and modeling", Springer series in statistics (Springer, New York ; London).

[61] Gallo, Julie Le, (2002), "Économétrie spatiale : L'autocorrélation spatiale dans les modèles de régression linéaire", *Economie & prévision* 155, 139-157.

[62] Gelfand, Alan E., Mark D. Ecker, John R. Knight, et C. F. Sirmans, (2004), "The dynamics of location in home price", *The Journal of Real Estate Finance and Economics* 29, 149-166.

[63] Genton, M., et D. Gorsich, (2002), "Nonparametric variogram and covariogram estimation with fourier–bessel matrices", *Computational Statistics & Data Analysis* 41, 47-57.

[64] Gillen, Kevin, Thomas Thibodeau, et Susan Wachter, (2001), "Anisotropic autocorrelation in house prices", *The Journal of Real Estate Finance and Economics* 23, 5-30.

[65] Goetzmann, William N., et Matthew Spiegel, (1997), "A spatial model of housing returns and neighborhood substitutability", *The Journal of Real Estate Finance and Economics* 14, 11-31.

[66] Grass, R., (1992), "The estimation of residential property values around transit station sites in washington, d.C", *Journal of Economics and Finance* 16, 139-146.

[67] Gratton, Yves, (2002), "Le krigeage : La méthode optimale d'interpolation spatiale", *Recherche* (Juin), 1-4.

[68] Haining, R., (1993). *Spatial data analysis in the social and environmental sciences* (Cambridge University Press).

[69] Haslett, John, (1997), "On the sample variogram and the sample autocovariance for non-stationary time series", *Journal of the Royal Statistical Society: Series D (The Statistician)* 46, 475-484.

[70] Hayunga, Darren, et R. Pace, (2010), "Spatial statistics applied to commercial real estate", *The Journal of Real Estate Finance and Economics* 41, 103-125.

[71] Hoesli, M., B. Thion, et C. Watkins, (1997), "A hedonic investigation of the rental value of apartments in central bordeaux", *Journal of Property Research* 14, 15-26.

[72] Hotelling, Harold, (1953), "New light on the correlation coefficient and its transforms", *Journal of the Royal Statistical Society. Series B (Methodological)* 15, 193-232.

[73] Isaaks, E., et R. Srivastava, (1988), "Spatial continuity measures for probabilistic and deterministic geostatistics", *Mathematical Geology* 20, 313-341.

[74] James, P. LeSage, (1998), "Econometrics: Matlab toolbox of econometrics functions", (Boston College Department of Economics).

[75] Jayet, Hubert, (1993). *Analyse spatiale quantitative une introduction* (Économica, Paris).

[76] Journel, André G., et Charles J. Huijbregts, (1991). *Mining geostatistics* (Academic press, London San Diego New York).

[77] Kain, John F., et John M. Quigley, (1970), "Measuring the value of housing quality", *Journal of the American Statistical Association* 65.

[78] Kelejian, Harry H., et Ingmar R. Prucha, (1998), "A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances", *The Journal of Real Estate Finance and Economics* 17, 99-121.

[79] Kelejian, Harry H., et Ingmar R. Prucha, (1999), "A generalized moments estimator for the autoregressive parameter in a spatial model", *International Economic Review* 40, 509-533.

[80] Kelejian, Harry H., et Dennis P. Robinson, (1993), "A suggested method of estimation for spatial interdependent models with autocorrelated errors, and an application to a county expenditure model", *Papers in Regional Science* 72, 297-312.

[81] Kent, John T., (1989), "Continuity properties for random fields", *The Annals of Probability* 17, 1432-1440.

[82] Kerry, R., et M. A. Oliver, (2007a), "Determining the effect of asymmetric data on the variogram. I. Underlying asymmetry", *Comput. Geosci.* 33, 1212-1232.

[83] Kerry, R., et M. A. Oliver, (2007b), "Determining the effect of asymmetric data on the variogram. Ii. Outliers", *Comput. Geosci.* 33, 1233-1260.

[84] Krige, Daniel, (1951), "A statistical approach to some basic mine valuation problems on the witwatersrand", *Journal of the Chemical, Metallurgical and Mining Society of South Africa* 52, 119-139.

[85] Kruskal, J., (1964), "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis", *Psychometrika* 29, 1-27.

[86] LeSage, James P., et R. Kelley Pace, (2004). *Spatial and spatiotemporal econometrics* (Elsevier JAI, Amsterdam ; London).

[87] LeSage, James P., et R. Kelley Pace, (2009). *Introduction to spatial econometrics* (CRC Press, Boca Raton ; London).

[88] Leuangthong, O., et C.V. Deutsch, (2005). *Geostatistics banff 2004* (Springer).

[89] Lin, Chu-Chia Steve, (1993), "The relationship between rents and prices of owner-occupied housing in taiwan", *The Journal of Real Estate Finance and Economics* 6, 25-54.

[90] Lin, Zhenguo, Eric Rosenblatt, et Vincent Yao, (2009), "Spillover effects of foreclosures on neighborhood property values", *The Journal of Real Estate Finance and Economics* 38, 387-407.

[91] Macpherson, David A., et G. Stacy Sirmans, (2001), "Neighborhood diversity and house-price appreciation", *The Journal of Real Estate Finance and Economics* 22, 81-97.

[92] Malpezzi, Stephen, (2008), "Hedonic pricing models: A selective and applied review", in *Housing economics and public policy* (Blackwell Science Ltd).

[93] Matheron, Georges, (1963), "Principles of geostatistics", *Economic Geology* 58, 1246-1266.

[94] Matheron, Georges, (1965). *Les variables régionalisées et leur estimation* (Masson et Cie, Paris).

[95] McCluskey, William J., et Richard A. Borst, (2011), "Detecting and validating residential housing submarkets: A geostatistical approach for use in mass appraisal", *International Journal of Housing Markets and Analysis* 4, 290-318.

[96] McDonald, J. F., et D. P. McMillen, (1990), "Employment subcenters and land values in a polycentric urban area: The case of chicago", *Environment and Planning A* 22, 1561-1574.

[97] McMillen, Daniel P., et John F. McDonald, (1998), "Suburban subcenters and employment density in metropolitan chicago", *Journal of Urban Economics* 43, 157-180.

[98] Melanie, Wall, (2004), "A close look at the spatial structure implied by the car and sar models", *Journal of Statistical Planning and Inference* 121, 311-324.

[99] Moran, P. A. P., (1948), "The interpretation of statistical maps", *Journal of the Royal Statistical Society. Series B (Methodological)* 10.

[100] Oden, Neal L., et Robert R. Sokal, (1986), "Directional autocorrelation: An extension of spatial correlograms to two dimensions", *Systematic Biology* 35, 608-617.

[101] Pace, R. Kelley, et Ronald Barry, (1997), "Quick computation of spatial autoregressive estimators", *Geographical Analysis* 29, 232-247.

[102] Pace, R. Kelley, Ronald Barry, John M. Clapp, et Mauricio Rodriquez, (1998), "Spatiotemporal autoregressive models of neighborhood effects", *The Journal of Real Estate Finance and Economics* 17, 15-33.

[103] Pace, R. Kelley, Ronald Barry, et C. F. Sirmans, (1998), "Spatial statistics and real estate", *The Journal of Real Estate Finance and Economics* 17, 5-13.

[104] Pace, R. Kelley, et Otis W. Gilley, (1997), "Using the spatial configuration of the data to improve estimation", *The Journal of Real Estate Finance and Economics* 14, 333-340.

[105] Pace, R. Kelley, et James P. LeSage, (2004), "Spatial statistics and real estate", *The Journal of Real Estate Finance and Economics* 29, 147-148.

[106] Paelinck, Jean H. P., et Leo H. Klaassen, (1979). *Spatial econometrics* (Saxon House, Farnborough).

[107] Pebesma, E., et C. G. Wesseling, (1998), "Gstat: A program for geostatistical modelling, prediction and simulation", *Computers & Geosciences* 24, 17-31.

[108] Quigley, J.M., (1994). *A simple hybrid model for estimating real estate price indexes* (Institute of Business and Economic Research, University of California at Berkeley, Center for Real Estate and Urban Economics).

[109] Rivoirard, Jacques, (1995). *Concepts et méthodes de la géostatistique*.

[110] Robin A, Dubin, (1992), "Spatial autocorrelation and neighborhood quality", *Regional Science and Urban Economics* 22, 433-452.

[111] Rodriguez, Mauricio, C. Sirmans, et Allen Marks, (1995), "Using geographic information systems to improve real estate analysis", *Journal of Real Estate Research* 10, 163-173.

[112] Rogers, William H., et William Winter, (2009), "The impact of foreclosures on neighboring housing sales", *Journal of Real Estate Research* 31.

[113] Rosen, S., (1974), "Hedonic prices and implicit markets: Product differentiation in pure competition", *Journal of Political Economy* 82, 34-null.

[114] Rosiers, Francois Des, Antonio Lagana, et Marius Theriault, (2001), "Size and proximity effects of primary schools on surrounding house values", *Journal of Property Research* 18, 149-168.

[115] Rossi, Richard E., David J. Mulla, Andre G. Journel, et H. Franz Eldon, (1992), "Geostatistical tools for modeling and interpreting ecological spatial dependence", *Ecological Monographs* 62, 277-314.

[116] Schuetz, Jenny, Vicki Been, et Ingrid Gould Ellen, (2008), "Neighborhood effects of concentrated mortgage foreclosures", *Journal of Housing Economics* 17, 306-319.

[117] Shimizu, Chihiro, et Kiyohiko Nishimura, (2007), "Pricing structure in tokyo metropolitan land markets and its structural changes: Pre-bubble, bubble, and post-bubble periods", *The Journal of Real Estate Finance and Economics* 35, 475-496.

[118] Simon, Arnaud, et Richard Malle, (2009). *Introduction à la finance et à l'économie de l'immobilier* (Économica, Paris).

[119] Strebelle, Sebastien, et Tuanfeng Zhang, (2005), "Non-stationary multiple-point geostatistical models geostatistics banff 2004", in Oy Leuangthong, et Clayton V. Deutsch, eds.: (Springer Netherlands).

[120] Sun, Hua, Yong Tu, et Shi-Ming Yu, (2005), "A spatio-temporal autoregressive model for multi-unit residential market analysis", *The Journal of Real Estate Finance and Economics* 31, 155-187.

[121] Thériault, M., F. Des Rosiers, P.Y. Villeneuve, et Y. Kestens, (2003), "Modelling interactions of location with specific value of housing attributes", *Property Management* 21.

[122] Topa, G., (1997). *Social interactions, local spillovers and unemployment* (New York University, Faculty of Arts and Science, Department of Economics).

[123] Tse, Raymond Y. C., (2002), "Estimating neighbourhood effects in house prices: Towards a new hedonic model approach", *Urban Studies* 39, 1165-1180.

[124] Tu, Yong, Hua Sun, et Shi-Ming Yu, (2007), "Spatial autocorrelations and urban housing market segmentation", *The Journal of Real Estate Finance and Economics* 34, 385-406.

[125] Tu, Yong, Shi-Ming Yu, et Hua Sun, (2004), "Transaction-based office price indexes: A spatiotemporal modeling approach", *Real Estate Economics* 32, 297-328.

[126] Valente, James;, ShanShan; Wu, Alan; Gelfand, et C.F. Sirmans, (2005), "Apartment rent prediction using spatial modeling", *Journal of Real Estate Research* 27.

[127] Whittle, P., (1954), "On stationary processes in the plane", *Biometrika* 41, 434-449.

[128] Wilhelmsson, Mats, (2000), "The impact of traffic noise on the values of single-family houses", *Journal of Environmental Planning and Management* 43, 799-815.

[129] Wilhelmsson, Mats, (2002), "Spatial models in real estate economics", *Housing, Theory and Society* 19, 92-101.

[130] Witte, Ann D., Howard J. Sumka, et Erekson Homer, (1979), "An estimate of a structural hedonic price model of the housing market: An application of rosen's theory of implicit markets", *Econometrica* 47.

[131] Won Kim, Chong, Tim T. Phipps, et Luc Anselin, (2003), "Measuring the benefits of air quality improvement: A spatial hedonic approach", *Journal of Environmental Economics and Management* 45, 24-39.

[132] Zimmerman, Dale, (1993), "Another look at anisotropy in geostatistics", *Mathematical Geology* 25, 453-470.

LISTE DES TABLEAUX

LISTE DES FIGURES

TABLE DES MATIERES

Résumé

La présence de dépendance spatiale des prix immobiliers impose aux méthodes d'estimation de prendre en compte cet élément. Les deux approches de la statistique spatiale sont l'économétrie spatiale et la géostatistique. La géostatistique estime directement la matrice de variance-covariance en supposant que la covariance entre les observations dépend inversement de la distance séparant leur localisation. L'économétrie spatiale définit et intègre la matrice d'interaction spatiale dans un modèle de régression hédonique. Si ces deux méthodes sont possibles pour étudier la dépendance spatiale des prix immobiliers dans des contextes variés, il n'existe cependant pas de règles très claires quant au choix de la méthode à sélectionner. Cette thèse procède à un examen détaillé de ces deux approches afin de pouvoir en distinguer les ressemblances et les différences, les avantages et les inconvénients. Des exemples d'application de chaque approche dans une étude immobilière sont présentés. La géostatistique est utilisée pour analyser la stationnarité du variogramme, ainsi que la sensibilité du variogramme aux paramètres de l'estimation hédonique. Il en ressort qu'il convient de distinguer l'influence spatiale sur le prix immobilier selon deux niveaux : l'effet de voisinage et l'effet de contiguïté. Le modèle d'économétrie spatiale est utilisé pour tenter d'identifier économétriquement le quartier dominant du marché immobilier d'une ville.

Mots-clés : Statistique spatiale, Géostatistique, Econométrie spatiale, Evaluation de biens immobiliers, Non-stationnarité spatiale, Quartier dominant

Abstract

Geostatistics and spatial econometrics are two spatial statistical approaches used to deal with spatial dependence. Geostatistics estimates directly the variance-covariance matrix by assuming that the covariance among observations depends inversely on the distance between their locations, called the covariogram. Spatial econometrics defines and integrates the spatial interaction matrix in a hedonic regression model. In real estate, price estimation should take into account these spatial characteristics because property prices are correlated. Hence, these two approaches are commonly used to study the spatial dependence of the real estate prices in many contexts. However, a definite rule in selection these statistic approaches has not been established. This thesis examined these two approaches in order to distinguish the similarities, differences, advantages, and disadvantages of each methodology. Some examples of their applications in a real estate study. The geostatistics is used to analyze the stationarity of the variogram and its sensitivity depending on the parameters added in hedonic estimation. The result showed that it is necessary to distinguish the spatial influence on real estate prices in two levels: neighborhood effect and contiguity effect. The spatial econometric is used to define econometrically the real estate market dominant area.

Key-words: Spatial statistics, Geostatistics, Spatial econometrics, Real estate price estimation, Non-stationary variogram, Dominant area