UNIVERSITÉ DE SHERBROOKE
École de gestion

Intégration de la dimension spatiale dans les modèles de prix hédoniques en évaluation immobilière

par
Hugo Lévesque

Mémoire présenté à l'École de gestion
en vue de l'obtention du grade de
M. Sc. en administration
Stratégie de l'intelligence d'affaires

Mai 2019

UNIVERSITÉ DE SHERBROOKE
École de gestion

Intégration de la dimension spatiale dans les modèles de prix hédoniques en évaluation immobilière

Hugo Lévesque

a été évalué par un jury composé des personnes suivantes :

Professeure Jessica Lévesque, directrice de recherche
Professeur Jean Cadieux, membre du jury
Professeure Jennifer Bélanger, membre du jury

Mémoire accepté le : _____

SOMMAIRE

Le présent mémoire aborde le thème de l'intégration de la dimension spatiale dans les modèles de prix hédoniques (MPH) en évaluation immobilière. Par une étude empirique, réalisée en contexte d'évaluation municipale à la ville de Laval, il comporte deux principaux objectifs : le premier étant de comparer les performances prédictives de modèles intégrant la dimension de la localisation de manière distincte, et le second étant de valider si le découpage *a priori* du territoire et la connaissance approfondie de ce dernier sont nécessaires pour optimiser la performance prédictive des modèles statistiques en évaluation immobilière.

Les approches testées préconisent les moindres carrés, ordinaires et pondérés, et omettent par conséquent de considérer certains modèles autorégressifs populaires, qui se fondent généralement sur le maximum de vraisemblance. Les cinq modèles évalués sont les suivants : un MPH sans découpage territorial (NAIF), un MPH avec découpage par des évaluateurs professionnels (EXPERT), les régressions géographiquement pondérées (RGP), un MPH appliquant une segmentation fuzzy (FUZZY), et finalement un MPH intégrant la procédure itérative de segmentation aléatoire (PISA), une technique novatrice développée dans le cadre du présent mémoire.

La performance des modèles est évaluée en fonction d'une technique de validation croisée sur les indicateurs de performance comportant 100 itérations, chacune calibrant les modèles sur la base d'un échantillon d'apprentissage, sélectionné aléatoirement, et évaluant la performance sur un échantillon de test, strictement constitué des données non utilisées pour calibrer ceux-ci.

Les modèles testés comportent la même spécification, à la seule distinction que quatre d'entre eux intègrent la dimension spatiale par une approche qui leur est

spécifique. Chacun est calibré sur la base du même jeu de données, comportant quelque 4 592 transactions *bona fide* de propriétés résidentielles de type unifamilial indivis, survenues entre le 1er janvier 2016 et le 1er juillet 2018, sur le territoire de la ville de Laval.

Les résultats obtenus, au terme de la présente étude, fournissent des conclusions fort intéressantes.

Dans un premier temps, les résultats générés par la technique de validation croisée favorisent le modèle PISA en termes de performance prédictive. Celui-ci présente un taux d'erreur moyen de 5,98% et estime 81,94% des prix avec un taux d'erreur en deçà de 10%. Également, il contrôle efficacement l'autocorrélation spatiale des résidus, tel qu'en témoigne la p-value de la statistique du I de Moran de 0,3822, qui excède largement le seuil critique de 5% fixé dans la présente étude. Nous notons aussi que, sur l'ensemble des ventes, ce modèle comporte un R^2 de 93,99%, qui se veut le plus élevé parmi les cinq modèles testés. Pour sa part, le modèle EXPERT, qui adopte une segmentation réalisée par des évaluateurs professionnels, a également généré d'excellents résultats, à la fois en termes de performance prédictive et de contrôle de l'autocorrélation spatiale des résidus. Néanmoins, nous notons que la PISA a non seulement surpassé, ne serait-ce que légèrement, la segmentation par experts au niveau de la performance prédictive, mais elle l'a fait en substituant les quelque 305 variables binaires, identifiant les unités de voisinage et les sous bassins d'analyse du modèle EXPERT, par une seule variable explicative : l'indice de prix hédonique de localisation (IPHL). Qui plus est, en calculant les taux d'erreur moyens générés par ces deux modèles, au sein de chaque unité de voisinage, nous constatons que la PISA ressort gagnante dans 52,11% d'entre elles, lesquelles englobent 58,97% des propriétés de l'inventaire.

Dans un second temps, une autre conclusion importante de ce mémoire est à l'effet que la performance prédictive respective des modèles PISA et EXPERT est

autocorrélée dans l'espace. Nous observons effectivement que le premier tend à surpasser le second dans les unités de voisinage où les forces spatiales, c'est-à-dire les aménités caractérisant un emplacement donné, se manifestent sur les prix d'une manière plus continue que discrète. À l'inverse, le modèle EXPERT ressort comme le plus précis dans les secteurs dominés par des forces discrètes. Nous concluons donc que ces deux approches se veulent complémentaires, et non concurrentes. La réconciliation de celles-ci, ou encore le choix optimal entre la PISA et le découpage *a priori* par experts, pour un secteur donné ou une propriété donnée, se veut une piste intéressante pour une recherche future. Or, sur la base des résultats obtenus, il ressort qu'un découpage *a priori* se veut l'un des deux ingrédients d'une segmentation optimale du territoire dans les MPH, à tout le moins sur le territoire de la ville de Laval. Bien que la piste qu'un tel découpage puisse être réalisé strictement sur la base des données demeure à explorer, nous voyons pour l'heure difficilement comment un algorithme pourrait se substituer efficacement au jugement de l'expert, dans les secteurs comportant peu de transactions et où interviennent diverses forces spatiales discrètes significatives. La pertinence du rôle de l'expert dans la segmentation territoriale n'est donc aucunement remise en cause au terme de la présente étude.

Dans un troisième temps, le présent mémoire fait ressortir l'importance de la localisation dans la formation des prix immobiliers résidentiels, sur le territoire de la ville de Laval. À cet effet, les résultats démontrent qu'une segmentation adéquate du territoire accroît significativement les performances prédictive et explicative des MPH. À titre d'exemple, la PISA fait passer le taux d'erreur moyen du modèle NAIF de 8,25% à 5,98%, ce qui constitue une amélioration de 27,52%. Également, sur l'ensemble des ventes, le R^2 de 93,99% du modèle PISA surpasse largement celui du modèle NAIF qui est de 86,91%. Également, en comparant certains estimateurs du modèle NAIF à ceux des quatre autres modèles, nous validons que l'omission de considérer les forces spatiales dans les MPH a pour effet de biaiser certains estimateurs, en leur attribuant de manière indue l'effet de ces forces spatiales omises. Qui plus est, nous observons que les estimateurs varient selon la segmentation territoriale retenue.

Cette constatation n'est sans doute pas étrangère à la problématique de l'aire spatiale modifiable. Il s'ensuit d'une importante réserve au niveau de l'interprétation des coefficients. À cet effet, nous suggérons que les effets *ceteris paribus*, mesurés par les MPH en évaluation immobilière, se doivent de toujours être interprétés par rapport à une spécification et une segmentation territoriale donnée, c'est-à-dire en compagnie de l'ensemble des autres coefficients, et non de manière isolée. Cette conclusion revêt toute son importance si les estimateurs des MPH sont destinés à ajuster les prix de vente des comparables, dans l'application de la technique des prix de vente ajustés.

Dans un quatrième temps, ce mémoire souligne l'importance d'évaluer la performance prédictive des MPH par un procédé prenant en compte le surapprentissage, que nous définissons comme la propension d'un modèle à surestimer sa performance prédictive, c'est-à-dire à générer une précision qui ne se matérialise que sur les observations utilisées pour le calibrer, et non sur des observations indépendantes du processus de calibration. Les résultats démontrent clairement que les indicateurs obtenus, sur la base de modèles calibrés sur l'ensemble des ventes, surestiment leur capacité respective à générer des estimations précises sur de nouvelles observations, totalement indépendantes du processus de calibration. Ce mécanisme a été simulé dans notre étude par la technique de validation croisée, ce qui a surtout fait ressortir une piètre généralisation de la part des RGP. À cet effet, nous suggérons deux pistes pour évaluer plus justement la performance prédictive des RGP : la première est de pondérer l'observation du sujet qui s'est transigé à 0 au lieu de 1 dans les régressions, et la seconde est de calculer le rayon optimal (« *bandwidth* ») sur la base d'un échantillon totalement indépendant des ventes utilisées pour mesurer la performance prédictive.

Pour conclure, l'originalité de la présente étude se décline sous plusieurs angles. D'une part, elle se démarque par son application pratique dans le domaine de l'évaluation municipale : les résultats obtenus guideront la conception du rôle d'évaluation triennal de 2019-2020-2021 à la ville de Laval. D'autre part, le présent

mémoire inaugure une toute nouvelle approche pour intégrer la dimension spatiale dans les MPH, la PISA, qui a généré la meilleure performance prédictive parmi les modèles testés. Cette nouvelle technique est la première à préconiser un découpage aléatoire du territoire, sur la stricte base de la proximité des propriétés. Elle s'appuie sur la première loi de la géographie, formulée par Tobler, à l'effet que les objets localisés à proximité interagissent davantage que les objets distants, et sur la prémisse qu'il existe non pas un seul, mais bien plusieurs découpages *a priori* admissibles. La procédure itérative, suggérée par la PISA, vient en quelque sorte répliquer le raisonnement d'un analyste qui souhaiterait obtenir, via un MPH, plusieurs estimations *a priori* de la valeur des emplacements, sur la base de découpages distincts du territoire regroupant des propriétés connexes ou voisines. Aussi, nous n'avons répertorié aucune autre étude en évaluation immobilière distinguant les propriétés à évaluer, c'est-à-dire l'inventaire, de celles ayant été transigées, soit les ventes. Une telle manière de procéder assure l'applicabilité des résultats à la population visée.

LISTE DES SIGLES

AS : Autocorrélation spatiale

IPHL : Indice de prix hédonique de localisation

MCO : Moindres carrés ordinaires

MCRL : Modèle classique de la régression linéaire

MPH : Modèle de prix hédoniques

PISA : Procédure itérative de segmentation aléatoire

RGP : Régressions géographiquement pondérées

REMERCIEMENTS

À 37 ans, et avec deux enfants en bas âge, un tel mémoire, n'aurait pu être possible sans deux éléments : un rêve et des gens qui m'ont permis de continuer d'y croire.

Ce rêve est sans doute né en 2006, d'une passion que j'ai développée pour l'approche hédonique et qui m'a été transmise par le professeur François Des Rosiers, de l'Université Laval. Au moment d'écrire ces lignes, je constate que cette passion ne se sera jamais estompée, malgré les années et les embuches. Ces dernières sont omniprésentes lorsqu'on souhaite innover. Sincères remerciements à vous, M. Des Rosiers, de m'avoir transmis votre passion et votre vision de l'évaluation immobilière, ainsi que pour vos conseils et votre générosité au fil des années. Vous êtes un professeur d'exception et avez fortement influencé mon parcours, non seulement professionnel, mais aussi personnel. Je vous en suis profondément reconnaissant.

En cours de route, j'ai rencontré plusieurs gens merveilleuses, étudiants et professeurs. Je tiens notamment à remercier l'équipe d'enseignants du département d'économique de l'Université de Sherbrooke, qui ont fait preuve d'une grande générosité, en m'offrant leur aide à un moment où il m'était rendu impossible de concilier les cours et ma vie personnelle. Un merci spécial à David Dupuis, François Delorme et au professeur Mario Fortin, pour son excellent cours d'introduction à l'économétrie. Je recommande fortement ce programme en économique pour sa qualité et celle de ses enseignants.

Ce périple n'aurait su être le même sans ma rencontre d'une autre personne d'exception, qui sera devenue ma directrice de mémoire, la professeure Jessica Lévesque. Si la simplification se veut l'ultime sophistication, Jessica est passée maître dans l'art d'expliquer et rendre accessibles, à des gens d'affaires, des techniques et

algorithmes autrement complexes. Un merci spécial à toi Jessica, d'abord pour la grande qualité de tes cours, qui m'ont permis non seulement de m'améliorer dans l'application des modèles linéaires, mais également d'ajouter plusieurs autres outils performants à mes connaissances. Tes enseignements ont façonné concrètement la manière dont on conçoit des modèles à la ville de Laval. Également, merci pour tes généreux conseils, ta disponibilité, ton ouverture d'esprit, ton intérêt marqué à l'endroit de mon projet de recherche, et pour avoir su identifier un sujet d'intérêt qui allait devenir l'un des thèmes principaux de mon mémoire : l'autocorrélation spatiale. J'ai vivement l'impression d'avoir beaucoup gagné en faisant ta rencontre. Je te remercie de croire en mes projets et d'y ajouter encore plus de sens.

Ce retour aux études a certes impliqué de nombreux sacrifices, notamment celui de renoncer à un salaire pour une durée d'un an, mais surtout celui de me distancer et me priver de ma famille à de multiples reprises. Un merci bien senti à mon amoureuse, Julie, d'avoir pris la relève avec nos deux amours pendant mes absences à Sherbrooke, et pour avoir accepté que je mette ce rêve de l'avant. Je t'aime et t'en suis très reconnaissant. Également, merci à mes parents, Jean-Paul et Linda, d'avoir été présents et impliqués tout au long de ce périple, et pour avoir minimisé l'impact de mes absences sur Julie, Thomas et Samuel.

Sincères remerciements aux professeurs Jennifer Bélanger et Jean Cadieux, qui ont généreusement accepté de lire et évaluer le présent mémoire.

Finalement, merci à la ville de Laval de m'avoir consenti les données requises pour mener à bien ce mémoire, et de croire en la valeur de ce projet.

TABLE DES MATIÈRES

LISTE DES FIGURES

INTRODUCTION

Il est généralement reconnu, dans le domaine de l'évaluation immobilière, que la localisation constitue un facteur déterminant, sinon l'un des plus importants, dans l'estimation de la valeur d'un bien immobilier. Or, chaque immeuble se distingue par un emplacement qui lui est propre, et se veut *a fortiori* unique. Au-delà de cette unicité de la localisation, les propriétés immobilières présentent une multitude de caractéristiques physiques pour le moins hétérogènes. Ensemble, les attributs physiques (aire du bâtiment, âge, aire du terrain, présence d'un garage, etc.) et les attributs de localisation (homogénéité du secteur, accessibilité, proximité des commodités, profil socio-économique du voisinage, etc.) d'une propriété lui procurent un degré d'utilité, de rareté et de désirabilité déterminant sa valeur perçue auprès d'acheteurs potentiels sur le marché immobilier. La capacité économique de ceux-ci à acquérir un panier de caractéristiques à un prix donné, c'est-à-dire leur pouvoir d'achat, vient inexorablement conditionner la demande pour celui-ci, et par conséquent sa valeur sur le marché immobilier.

L'évaluateur professionnel est donc confronté à plusieurs défis lorsqu'il estime la valeur foncière d'une propriété. Un lien naturel évident existe entre la théorie hédonique (Rosen, 1974) et les défis rencontrés en évaluation immobilière. En effet, selon la théorie hédonique, sous certaines conditions, le prix d'un bien complexe (maison, voiture, etc.) peut être décomposé en une série de contributions marginales de ses divers attributs. Les prix implicites de ces derniers se veulent essentiellement tributaires de la propension des acheteurs à payer pour ceux-ci. Les modèles de prix hédoniques (MPH) combinent la théorie hédonique à la robustesse des moindres carrés ordinaires (MCO), et plus spécifiquement du modèle de régression linéaire multiple. Les MPH sont largement utilisés par les évaluateurs et chercheurs en évaluation depuis bon nombre d'années, notamment pour analyser la dynamique du marché immobilier, dégager des valeurs contributives, et estimer des prix de vente.

Or, à l'instar de tout modèle statistique linéaire fondé sur les MCO, les MPH sont sujets à diverses problématiques d'ordre économétrique, à savoir la multicolinéarité, l'hétéroscédasticité et l'autocorrélation spatiale (AS) des résidus (Thériault et al., 2011, p. 236 et 237).

La multicolinéarité désigne une situation où une variable explicative se veut corrélée linéairement avec une ou plusieurs autres variables explicatives du modèle. À titre d'exemple, dans le domaine de l'immobilier, on peut s'attendre à ce que les maisons de plus grande superficie habitable comportent, en moyenne, davantage de pièces, de chambres à coucher et de salles de bain que les maisons plus petites. L'inclusion simultanée de ces quatre variables explicatives dans un MPH aurait donc pour effet attendu d'induire une certaine multicolinéarité au sein de celui-ci. Il est à noter que la multicolinéarité excessive ne biaise pas les prévisions, mais a pour effet d'invalider les cotes-t et les intervalles de confiance des estimateurs (Gujarati, 2004, p. 348 à 355).

Pour sa part, l'hétéroscédasticité dénote un cas où, pour une ou plusieurs variables explicatives du modèle, les variances des termes d'erreur ne sont pas constantes, pour l'ensemble des valeurs prises par celles-ci. À titre d'exemple, dans les MPH en immobilier, il n'est pas rare de constater que la dispersion des résidus est accrue pour les propriétés les plus luxueuses, en comparaison aux propriétés plus économiques. En pratique, on constate qu'une telle hétéroscédasticité est fréquemment induite par une spécification déficiente, c'est-à-dire par l'omission de variables explicatives importantes, la conjugaison de propriétés trop disparates au sein d'un même MPH, ou encore l'emploi d'une forme fonctionnelle inadaptée au contexte. Or, à moins d'être symptomatique d'une erreur de spécification, l'hétéroscédasticité ne biaise pas les prévisions (Gujarati, 2004, p. 393 et 394). Toutefois, celle-ci a pour effet

d'invalider les variances des estimateurs, les cotes-t, les statistiques F et les intervalles de confiance des estimateurs (Gujarati, 2004, p. 398 à 400).

Quant à l'AS des résidus, qui est également connue sous le nom de dépendance spatiale, elle réfère à une situation où les résidus d'une régression tendent à être corrélés dans l'espace, c'est-à-dire à être entourés soit par des résidus de même signe (AS positive), soit par des résidus de signe opposé (AS négative). En présence d'AS significative des résidus, les termes d'erreur de la régression sont présumés ne pas être indépendants, ce qui va à l'encontre d'une hypothèse fondamentale du modèle classique de la régression linéaire (MCRL), selon laquelle les termes d'erreur se doivent d'être non corrélés entre eux (Gujarati, 2004, p. 70). L'AS des résidus a pour principales conséquences d'invalider les tests statistiques usuels et les intervalles de confiance des estimateurs, mais elle peut aussi biaiser les estimations si des facteurs importants de localisation sont omis (Thériault et al., 2011, p. 237).

Rappelons également qu'à la base, les hypothèses au soutien du MCRL présupposent une bonne spécification du modèle (Gujarati, 2004, p. 66 et 73 à 75). En effet, le modèle spécifié doit être le « vrai modèle » : il doit être linéaire dans les paramètres et dépourvu de biais de spécification, d'erreurs et de biais d'omission. C'est donc dire que les variables omises doivent être marginales en rapport à la variable dépendante d'intérêt et ne doivent pas être corrélées à la fois avec les résidus et l'une ou l'autre des variables explicatives du modèle (Stock et al., 2012, p. 82 à 84). Qui plus est, une mauvaise spécification biaise les estimations (Gujarati, 2004, p. 547 et 548).

Or, sur la base des précédentes discussions, il en ressort dès lors qu'une segmentation territoriale adéquate des sous marchés revêt une importance cruciale dans les MPH en évaluation immobilière, y compris ceux dont la vocation est de prédire les prix de vente. Mais cela présuppose-t-il que cette segmentation doit être réalisée *a priori* et sur la base de connaissances préalables quant aux forces du marché régissant le territoire sous étude?

Le présent mémoire aborde le thème de l'intégration de la dimension spatiale dans les MPH, sous un angle qui se veut plus pragmatique que théorique. Les appellations segmentation du territoire et découpage du territoire sont utilisées comme synonymes. Celles-ci réfèrent essentiellement aux diverses techniques permettant d'intégrer la dimension de la localisation dans les MPH, c'est-à-dire mesurer la valeur intrinsèque, ou encore les prix hédoniques, des emplacements.

Sur la base d'une étude empirique, réalisée en contexte d'évaluation municipale au sein de la ville de Laval, le premier objectif visé par cette étude est de comparer les performances prédictives de diverses segmentations du territoire dans les MPH, en vue d'une utilisation pratique dans le cadre de la conception du rôle d'évaluation triennal de 2019-2020-2021. Le second objectif de ce mémoire est de valider si l'intervention d'experts, à savoir le découpage *a priori* du territoire sur la base de connaissances préalables quant aux forces du marché s'y appliquant, est indispensable pour optimiser la performance prédictive des MPH. En résumé, cette étude tente de répondre aux questions qui suivent :

Q1 : Parmi les diverses segmentations du territoire testées, laquelle procure la meilleure performance prédictive?

Q2 : Une segmentation du territoire centrée sur les données peut-elle procurer une meilleure performance prédictive qu'un découpage *a priori* réalisé par des experts?

Il existe plus d'une technique permettant d'intégrer la dimension de la localisation dans les MPH et ne requérant ni découpage *a priori* du territoire ni connaissances approfondies sur celui-ci. De telles techniques laissent généralement les données estimer les prix hédoniques des emplacements, sans nécessiter formellement l'intervention d'experts. Trois approches centrées sur les données sont présentées et testées dans cette étude. Celles-ci reposent exclusivement sur les moindres carrés,

ordinaires ou pondérés, excluant d'emblée les modèles autorégressifs spatiaux, qui préconisent généralement le maximum de vraisemblance. Plus spécifiquement, nous abordons dans ce mémoire les régressions géographiquement pondérées (RGP), la segmentation *fuzzy* basée sur la proximité des propriétés et la procédure itérative de segmentation aléatoire (PISA). Ces trois approches sont ultimement comparées entre elles, ainsi qu'à une segmentation *a priori* réalisée par des évaluateurs professionnels. Nous procédons à cette comparaison en implantant successivement chaque segmentation au sein d'un MPH n'intégrant pas la dimension de la localisation. Celui-ci conserve la même spécification, c'est-à-dire la même variable dépendante et les mêmes 29 variables explicatives, à l'exception de la dimension de la localisation, qui est spécifique à chaque segmentation testée.

Or, un élément crucial de ce mémoire se veut la comparaison de divers modèles sur la base de leur performance prédictive, c'est-à-dire leur capacité respective à produire des estimations précises et non biaisées sur des données n'ayant pas servi à les calibrer. À cet effet, il convient de paraphraser l'économiste Milton Friedman, à l'effet que le but ultime de tout modèle est de produire des prévisions non triviales et valides sur des données qui n'ont pas encore été observées[1]. La propension d'un modèle à surestimer sa performance prédictive, c'est-à-dire à générer une précision qui ne se matérialise que sur les observations utilisées pour le calibrer, et non sur des observations indépendantes du processus de calibration, porte le nom de surapprentissage. Or, pour éviter d'avantager un modèle qui serait davantage sujet au surapprentissage, nous calculons les indicateurs de performance selon une technique de validation croisée sur les indicateurs de performance (Anselin, 1988, p. 249).

Voilà qui termine la présente introduction. La suite de ce mémoire s'annonce comme suit. Le premier chapitre aborde le cadre conceptuel de notre étude. Nous y présentons une revue de la littérature des principaux sujets d'intérêt, et les concepts

[1] Milton Friedman. « The Methodology of Positive Economics », dans *Essays in Positive Economics*, University of Chicago Press, Chicago, 1953. p. 7.

clés constituant le cadre *a priori* de ce mémoire. Le second chapitre traite de la méthodologie. Nous y expliquons, de manière détaillée, comment nous adaptons les concepts théoriques présentés, au contexte de notre étude. Le troisième chapitre présente et analyse les résultats. Finalement, le quatrième chapitre aborde certaines discussions en regard à l'étude (résultats, apport, forces et limites) et adresse certaines pistes pour des recherches futures.

CHAPITRE 1 : CADRE CONCEPTUEL

Jusqu'à maintenant, nous avons souligné l'importance du thème de l'intégration de la dimension spatiale dans les MPH en évaluation immobilière. Nous avons fait valoir qu'une segmentation territoriale adéquate est indispensable pour assurer la validité de l'inférence statistique et des estimations issues de tels modèles. Nous nous questionnons à savoir si une segmentation centrée sur les données peut se substituer efficacement à une segmentation *a priori* réalisée par des experts. Or, le présent chapitre aborde une revue de la littérature des principaux sujets d'intérêt, et présente les concepts clés constituant le cadre *a priori* de notre étude.

1.1 Revue de la littérature

1.1.1 Problématiques inhérentes à l'économétrie spatiale

Pour débuter cette revue de la littérature, il convient de présenter les principales problématiques rencontrées dans le cadre d'analyses économétriques intégrant la dimension spatiale.

1.1.1.1 Dépendance et hétérogénéité spatiale

Selon Anselin (1988, p. 1), l'économétrie traditionnelle omet de considérer deux problématiques importantes propres aux données spatiales : la dépendance spatiale et l'hétérogénéité spatiale. Cette lacune a entraîné le développement d'un tout nouveau champ d'expertise : l'économétrie spatiale.

La dépendance spatiale, qui se veut un synonyme de l'AS des résidus, réfère à la tendance des résidus à être corrélés dans l'espace en fonction de leur proximité. Celle-ci a pour effet de rendre les estimateurs des MCO inefficients et d'invalider leur

erreur standard (Thériault et al. 2011, p. 237). L'AS des résidus peut aussi être symptomatique de problèmes plus graves, tels une mauvaise spécification (Le Gallo, 2002). En pareils cas, un modèle sera biaisé, s'il omet de considérer des facteurs de localisation importants, et s'il attribue certaines variations inhérentes à ces facteurs omis à un ou plusieurs coefficients du modèle (Thériault et al., 2011, p. 215). Dans les MPH classiques, la prévention de l'AS des résidus passe par un bon découpage géographique *a priori* des unités de voisinage, mais un tel découpage est virtuellement assuré de contenir des erreurs de mesure pour deux raisons : premièrement les voisinages ne sont pas concrètement observables, et deuxièmement leur limite est inconnue sur une base *a priori* (Dubin, 1998).

Pour sa part, l'hétérogénéité spatiale fait référence à la variabilité des relations dans l'espace, qui engendre une non-stationnarité dans les paramètres. Pour bien introduire et illustrer ce concept, il convient de reprendre un exemple formulé par Fotheringham et al. (2002, p. 1), se rapportant à diverses données sur le climat aux États-Unis. À cet effet, les auteurs suggèrent d'imaginer un livre abordant ce sujet, et présentant exclusivement des statistiques agrégées sur l'ensemble du pays, telles les précipitations moyennes et le nombre d'heures d'ensoleillement moyen sur l'ensemble du territoire des États-Unis. Force est d'admettre qu'un tel livre serait pratiquement d'aucune utilité pour qualifier le climat d'une région en particulier, étant donné les écarts importants constatés parmi les diverses régions. Le fait de considérer une relation comme stationnaire dans un MPH, alors qu'elle ne l'est pas, limite à la fois le pouvoir explicatif et prédictif, et biaise les estimations (Bitter et al., 2007). Également, il engendre un important problème au niveau de l'interprétation des coefficients (Fotheringham et al., 1998).

Diverses solutions ont été développées pour aborder les problématiques de dépendance et d'hétérogénéité spatiale. Parmi elles figurent les régressions géographiquement pondérées (Fotheringham et al., 1998) et la méthode d'expansion spatiale, qui consiste à instaurer des variables interactives dans les MPH (e.g. Kestens

et al., 2006). Toutefois, cette dernière solution requiert une spécification *a priori* de la forme des équations d'expansion (Fotheringham et al., 2002, p. 16 et 17).

1.1.1.2 Nombre limité d'observations et problème des paramètres incidents (*« incidental parameter problem »*)

Par ailleurs, une autre problématique inhérente à l'économétrie spatiale se rapporte au nombre limité d'observations disponibles pour modéliser les phénomènes spatiaux, qui se veulent par essence nombreux et souvent complexes. À cet effet, il convient de noter que les forces spatiales se composent non seulement de facteurs exogènes observables (profil socio-économique, accessibilité, proximité des commodités, nuisances sonores ou visuelles, etc.), mais également de facteurs endogènes subtils, qui impliquent à la fois le sujet et les propriétés environnantes, et qui affectent les prix hédoniques des emplacements via le principe de la conformité (Des Rosiers et al., 2011). Selon ce principe, la valeur d'une propriété immobilière se veut influencée par son degré de similitude avec les autres propriétés environnantes.

Or, dans les MPH classiques, pour modéliser ces phénomènes, l'inclusion de nombreuses variables binaires identifiant les voisinages et, à plus forte raison, de variables interactives, consomment une importante quantité de degrés de liberté, limitant du même coup la capacité à estimer certains paramètres de manière fiable. Cette problématique, qui est loin d'être exclusive à l'évaluation immobilière, est connue sous le nom de problème des paramètres incidents (Neyman et Scott, 1948). L'analyste en évaluation immobilière est donc souvent confronté, en pratique, à un important dilemme, qui se résume soit à minimiser le biais induit par l'omission de certaines variables de localisation pertinentes, soit à minimiser la variance des estimateurs (Bourassa et al., 2003).

1.1.1.3 Aire spatiale modifiable

Finalement, il importe de citer, à titre d'enjeu propre à l'économétrie spatiale, la problématique de l'aire spatiale modifiable (Gehlke et Biehl, 1934). Selon cette dernière, les résultats analytiques sont intimement liés à la définition des unités spatiales, c'est-à-dire à leur niveau d'agrégation (« *scale effect* ») et au découpage choisi (« *zoning effect* ») (Fotheringham et al., 1991). Puisque les voisinages sont inconnus sur une base *a priori* et que leurs limites ne sont pas concrètement observables, et qu'en théorie il puisse y avoir autant d'unités spatiales à modéliser qu'il y a d'emplacements à évaluer, cette problématique se veut insoluble en pratique. En conséquence, l'analyste en évaluation immobilière doit relâcher une hypothèse fondamentale du MCRL, à savoir la quête de la spécification du vrai modèle, pour s'adonner à la poursuite de modèles incomplets mais utiles. Au sein de tels modèles, les estimateurs se doivent d'être interprétés conjointement et contextuellement à la spécification retenue, principalement en raison des liens entre les variables explicatives du modèle. Pour ainsi dire, les estimateurs générés par les MPH en évaluation immobilière mesurent l'effet *ceteris paribus* de leur variable associée, mais cet effet est relié à une spécification et une segmentation territoriale donnée, et non à la réalité absolue. Dans le présent mémoire, l'expression *ceteris paribus* est néanmoins utilisée, mais son utilisation demeure sujette à la réserve qui précède.

1.1.2 Découpage *a priori* du territoire dans les MPH

Nous avons évoqué, à la sous-section 1.1.1.1, que la prévention de la dépendance spatiale, dans les MPH classiques, passe généralement par un bon découpage géographique *a priori* des unités de voisinage. Il n'est donc pas surprenant que la littérature soit particulièrement abondante en ce qui a trait à ce sujet. Or, il ressort de celle-ci trois principales constatations : la première est qu'il existe une grande variété de critères pour découper le territoire sur une base *a priori*, la seconde est que la plupart des découpages du territoire améliorent la performance prédictive par rapport

à un MPH sans segmentation territoriale, et la troisième est qu'aucun critère de découpage ne fait formellement l'unanimité. Nous présentons les articles consultés, par ordre chronologique quant à leur année de publication.

Tout d'abord, dès 1970, Kain et al. ont appliqué une analyse factorielle sur 39 indices de qualité du voisinage obtenus par sondage. Les cinq facteurs issus de cette analyse ont ensuite été intégrés à titre de variables explicatives dans des régressions linéaires multiples. En guise de conclusion, les auteurs ont souligné l'importance de la qualité du voisinage dans la formation des prix immobiliers, une importance qualifiée d'au moins aussi importante que certaines variables quantitatives clés, telles que le nombre de chambres, le nombre de salles de bains et la superficie du terrain. Cette étude a donc validé le rôle prépondérant de la localisation dans la formation des prix immobiliers.

En 1981, Goodman a préconisé les limites administratives au sein d'une région métropolitaine de recensement afin d'évaluer la variabilité des coefficients d'un MPH selon le niveau d'agrégation choisi. Les résultats ont démontré que les prix implicites des services et commodités varient substantiellement dans le temps et selon la région administrative. Cette étude a donc validé l'existence de relations non stationnaires en évaluation immobilière, à la fois dans l'espace et dans le temps.

En 1990, Dubin et Sung ont quant à eux analysé les préférences des ménages à l'endroit de diverses caractéristiques de voisinage à l'aide du test statistique non imbriqué J, destiné à tester deux hypothèses concurrentes en présente de forte multicolinéarité ou de biais d'omission. Les auteurs ont statué que les variables de voisinage utilisées dans le cadre d'études hédoniques peuvent être classées en trois catégories : la race, le profil socio-économique et les services. Leurs résultats ont démontré que les caractéristiques du voisinage, c'est-à-dire la race et le profil socio-économique, sont importantes pour expliquer les préférences des ménages, tandis que les services ressortent comme relativement peu importants. Avec respect pour cette

étude, nous émettons néanmoins une réserve quant à sa généralisation à diverses villes québécoises contemporaines. À cet effet, il convient de reprendre les conclusions de Goodman (1981), à l'effet que les prix implicites des services et commodités varient à la fois dans le temps et dans l'espace.

En 1996, Maclennan et Tu ont testé l'analyse factorielle et en composantes principales sur la base de caractéristiques clés propres aux logements individuels. Celles-ci réfèrent à trois catégories distinctes : les variables internes propres aux logements, les variables de voisinages et les variables de proximité. Les facteurs issus de cette analyse ont par la suite été inclus dans un MPH pour en dériver des conclusions. Les auteurs ont conclu que les sous marchés existent en pratique, que ces derniers sont importants et qu'ils fluctuent dans le temps.

En 1998, Goodman et Thibodeau ont préconisé les modèles linéaires hiérarchiques qui permettent aux coefficients de varier en fonction des sous marchés. Ils ont conclu que la qualité des écoles primaires, telle que mesurée par des tests standardisés, est déterminante de la segmentation des sous marchés résidentiels dans la région métropolitaine de Dallas.

En 1999, Bourassa et al. ont utilisé les analyses en composantes principales et l'analyse de segmentation de type hiérarchique et par nuées dynamiques sur la base de critères des zones administratives et des logements individuels, afin de produire des segments fondés sur le score de chaque facteur extrait. La classification par nuées dynamiques est la seule ayant généré un gain significatif dans le MPH, en termes de performance prédictive pour la ville de Melbourne, par rapport à la segmentation *a priori*. Cette étude constitue un exemple à l'effet qu'une approche centrée sur les données peut, dans certains cas, générer une meilleure performance prédictive qu'un découpage *a priori*.

Par ailleurs, en 2003, Bourassa et al. ont conclu que le découpage territorial réalisé par des évaluateurs professionnels engendrait les meilleures performances prédictives dans les MPH. Cette conclusion est pour le moins intéressante, du point de vue de la présente étude, puisqu'un tel découpage constitue justement la segmentation territoriale que nous tentons de surpasser par une approche centrée sur les données.

En 2006, Clapp et al. ont proposé une méthode statistique visant à optimiser l'homogénéité, la contiguïté et la parcimonie des segments de voisinage créés. Ils ont entre autres utilisé les résidus d'un MPH en tant qu'indicateur de la valeur de la localisation, puis ont intégré ceux-ci dans un arbre de décision de type CART. Les résultats ont démontré que cette segmentation améliore la précision des estimations par rapport à un MPH sans segmentation.

Dans une perspective similaire, en 2007, Tu et al. ont proposé une méthode de segmentation visant à laisser les données découper les segments du marché immobilier. La technique proposée se fonde sur la structure de l'autocorrélation spatiale dans les résidus. Ces derniers déterminent ensuite des segments sur la base de leur corrélation. Les résultats ont eux aussi démontré qu'une telle segmentation améliore significativement la précision des estimations par rapport à un MPH sans segmentation.

En 2009, Lockwood a utilisé une analyse en composantes principales, sur la base de variables structurelles et spatiales, pour produire des facteurs qui ont par la suite été intégrés à un modèle de régressions géographiquement pondérées. Toutefois, les résultats n'ont pas été comparés à un MPH sans segmentation.

Par ailleurs, en 2010, Voisin et al. ont comparé une segmentation historico-morphologique à deux découpages administratifs, avant et après les fusions municipales. Leur conclusion est à l'effet que l'approche historico-morphologique se veut plus efficace que les découpages administratifs, tout en demeurant perfectible.

Finalement, en 2013, Helbich et al. ont préconisé une approche centrée sur les données de manière à générer des segments spatiaux homogènes et contigus. Pour ce faire, les auteurs ont utilisé des régressions semi locales, l'analyse en composantes principales, et l'algorithme SKATER qui se fonde sur des arbres couvrants minimaux de graphes. Leur conclusion est à l'effet qu'une telle segmentation optimise la performance prédictive dans les MPH par rapport à une segmentation fondée sur les unités administratives, ou encore générée exclusivement par les nuées dynamiques.

Pour conclure cette revue de la littérature sur le découpage *a priori* du territoire dans les MPH, nous soulignons qu'aucune étude relevée n'a envisagé la piste du découpage aléatoire de propriétés connexes ou voisines comme unique critère de segmentation dans les nuées dynamiques. Qui plus est, aucune n'a préconisé un processus itératif permettant à un MPH de dégager, sur une base *a priori*, non pas une seule, mais bien plusieurs appréciations de la valeur d'un emplacement. Ces concepts sont à la base de la procédure itérative de segmentation aléatoire, introduite dans le cadre du présent mémoire.

1.1.3 Régressions géographiquement pondérées (RGP)

Les RGP constituent des régressions au sein desquelles les observations sont pondérées selon une fonction inverse de leur distance géographique les séparant de chaque emplacement pour lequel on désire estimer une régression. La présente sous-section dresse une revue de la littérature de leur utilisation dans diverses études à des fins prédictives en évaluation immobilière. Leur fonctionnement détaillé est abordé à la sous-section 1.2.2.

Tout d'abord, Farber et Yeates (2006) ont comparé, pour le marché immobilier de Toronto, les performances prédictives des RGP à un MPH traditionnel, un modèle SAR basé sur les MCO et un modèle préconisant la technique de « *moving window regression* », qui se veut un cas spécial des RGP. Les auteurs ont conclu que les RGP

surpassent les autres approches, tant au niveau explicatif que prédictif, ainsi que pour minimiser l'autocorrélation spatiale des résidus. Néanmoins, ils ont dénoté la présence de certains coefficients jugés contraires à la logique, majoritairement concentrés dans quelques voisinages, lesquels n'affectent toutefois pas la fiabilité des estimations.

Des résultats similaires ont été obtenus par Bitter et al. (2007), qui comparaient les RGP à un MPH classique et à deux modèles appliquant la méthode d'expansion spatiale, dont l'un incluait un terme autorégressif de la variable dépendante, pour le marché immobilier de Tucson, en Arizona. Les résultats ont démontré la supériorité des RGP en termes de pouvoir prédictif et explicatif. Néanmoins, les auteurs ont eux aussi soulevé l'existence de coefficients jugés contraires à la logique.

À cet effet, des études ont démontré que les coefficients jugés contraires à la logique, rapportés dans les études portant sur les RGP, seraient en faits induits par une multicolinéarité excessive générée par celles-ci, en raison de leur mécanisme d'action qui réutilise systématiquement plusieurs fois les mêmes observations dans un voisinage donné (Thériault et al., 2011, p. 222), lesquelles observations comportent souvent des caractéristiques très fortement corrélées entre elles.

Par ailleurs, Gao et al. (2006) ont testé les performances prédictives de quatre modèles distincts : un modèle de base, un modèle de dépendance spatiale intégrant les variables explicatives décalées spatialement (*« spatial lag »*), les RGP et un modèle mixte combinant à la fois les RGP et les variables explicatives décalées spatialement. L'étude a été réalisée sur la base d'un échantillon de 190 ventes de résidences détachées à Tokyo. Les résultats ont pointé en faveur du modèle mixte, puis des RGP, en termes de taux d'erreur moyen en valeur absolue, d'erreur moyenne en yens et de la somme du carré des résidus.

Également, Páez et al. (2008) ont comparé, pour le marché de Toronto, trois modèles spatiaux à un MPH sans découpage territorial, appelé le modèle naïf. Les trois

modèles spatiaux sont : la « *moving window regression* », les RGP et le kriging. Les résultats indiquent que les trois modèles spatiaux surpassent le modèle naïf en termes de performance prédictive. Toutefois, cette étude a fait ressortir que les RGP et la « *moving window regression* » ont généré les meilleures performances prédictives parmi les approches testées.

Dans une perspective similaire, Bidanset et al. (2014) ont comparé, pour le marché des résidences unifamiliales de Norfolk en Virginie, la performance de trois modèles, à savoir : les RGP, un modèle sans segmentation territoriale et un modèle autorégressif spatial intégrant la variable dépendante décalée. Les auteurs ont conclu que les RGP, de manière globale, surpassent les deux autres modèles en ce qui a trait à minimiser les coefficients de variation des estimations et l'AIC. Toutefois, ils ont fait ressortir que la performance des trois modèles a varié selon les voisinages, et qu'aucun d'eux ne s'est avéré systématiquement meilleur sur l'ensemble du territoire. Ils ont souligné l'importance de tester divers modèles au sein des divers voisinages, de manière à sélectionner le plus approprié en un endroit donné.

Finalement, en 2016, Helbich et al. ont testé, dans le marché immobilier résidentiel de Vienna en Autriche, la méthode d'expansion spatiale, la « moving window régression », les RGP et la technique de « eigenvector spatial filtering ». Au niveau de la performance prédictive, les résultats obtenus sur la base de 100 échantillons indépendants du processus de calibration des modèles ont eux aussi avantagé les RGP.

En résumé, il appert que les RGP s'avèrent particulièrement performantes au niveau prédictif et pour réduire l'AS des résidus. On leur reproche principalement de générer certains coefficients contraires à la logique en des lieux spécifiques, ce qui n'est pas étranger à la forte multicolinéarité induite par les régressions locales, et n'est théoriquement d'aucune incidence sur la précision des prévisions.

1.1.4 La segmentation *fuzzy*

Les méthodes conventionnelles de segmentation (« *hard clustering* ») requièrent que chaque observation d'une solution appartienne à un seul segment. Or, la théorie des ensembles *fuzzy* (Zadeh, 1965) est venue proposer un tout nouveau paradigme, à l'effet qu'il existe une incertitude quant à l'appartenance, laquelle a été décrite par l'auteur comme une fonction d'appartenance. L'application de cette théorie à l'analyse en segmentation a été instaurée dès 1966 par Bellman, Kalaba et Zadeh, ce qui donnera naissance à la segmentation *« soft »*, également appelée segmentation *fuzzy*. Cette dernière permet à chaque observation d'appartenir à plus d'un segment, et ce, avec un coefficient d'appartenance associé.

Si les modèles et systèmes *fuzzy*, inspirés de la théorie des ensembles *fuzzy*, ont été passablement abordés dans la littérature en évaluation immobilière, tel n'est pas le cas de la segmentation *fuzzy*, qui a fait l'objet d'une attention discrète. Pourtant, dans leurs études, Hwang et Thill (2007, 2009) ont fait ressortir que la segmentation *fuzzy* a surpassé la segmentation *« hard »* pour minimiser l'erreur prédictive dans les MPH, et a réduit l'erreur d'estimation de l'ordre de 34.50% par rapport à un MPH sans segmentation. Il appert donc qu'un tel algorithme mérite d'être testé dans le cadre de notre étude.

1.2 Concepts clés constituant le cadre a priori de notre étude

La revue de la littérature fait clairement ressortir que la théorie hédonique, et à plus forte raison, les modèles de prix hédoniques (MPH), sont largement utilisés dans les études en évaluation immobilière, et ce, depuis son avènement en 1974. Or, ce mémoire ne fait pas exception à la règle, et situe l'approche hédonique au cœur de ses concepts clés.

1.2.1 L'approche hédonique et les MPH en évaluation immobilière

En introduction, nous avons déjà référé à la théorie hédonique, selon laquelle le prix d'un bien hétérogène peut être décomposé en une série de prix implicites de ses divers attributs. Dans un MPH, le prix de vente, *Y*, s'exprime comme une fonction de l'ensemble des caractéristiques observables d'une propriété, c'est-à-dire :

$$Y = X\beta + \mu \qquad \text{(équation 1)}$$

où β est un vecteur contenant les prix hédoniques de chacune des caractéristiques observables X et μ constitue le terme d'erreur aléatoire identiquement et indépendamment distribué.

Par ailleurs, outre les caractéristiques physiques observables des propriétés immobilières, nous avons souligné que chacune d'elles comporte un emplacement unique, constitué de caractéristiques à la fois objectives et subjectives. La distinction entre ces deux types est plus complexe qu'elle n'y paraît.

Pour illustrer ce propos, nous relatons l'histoire, entièrement fictive, de Jean et Janice, les parents de deux jeunes enfants. Ceux-ci convoitent une maison à vendre, localisée à proximité d'une école primaire de quartier très réputée, au sein de laquelle ils aimeraient beaucoup inscrire leurs enfants. Jean et Janice entreprennent donc leurs recherches, pour valider si cette maison correspond véritablement à celle de leurs rêves. Sur la base d'une carte aérienne, Jean estime que la maison en question se situe à une distance euclidienne de seulement 127 mètres de l'école tant convoitée. Il empresse donc Janice de se rendre sur place, afin de pouvoir mieux apprécier le secteur. Arrivés à destination, Jean et Janice arpentent les lieux, sans néanmoins éprouver la même excitation qui les y avait emmenés. Ils constatent rapidement que, malgré la proximité de l'école, à vol d'oiseau, le trajet le plus court pour s'y rendre, à pieds, requiert une bonne dizaine de minutes. Dommage que le secteur ne soit pas muni de sentiers

piétonniers, pour éviter un tel détour. Qui plus est, en continuant leur visite du secteur, Jean et Janice sont pour le moins déçus de constater que les routes à emprunter, pour se rendre à l'école, ne sont munies d'aucun trottoir, et que celles-ci sont étroites et surchargées de voitures. Ils craignent pour la sécurité de leurs enfants. À cet effet, la limite de vitesse de 50 km/h leur apparaît excessive, compte tenu des facteurs qu'ils ont identifiés. Or, comble du hasard, en circulant devant la maison convoitée, ils croisent le propriétaire, avec qui ils entament une conversation. Celui-ci leur apprend, au grand désarroi de Jean et Janice, que ses propres enfants n'ont jamais réussi à s'inscrire à cette école primaire, puisque leur maison est localisée tout juste en dehors de la zone définie par la commission scolaire de son quartier, pour permettre une inscription à cette école.

Cet exemple sert à démontrer que l'objectivité apparente d'une mesure, dans cet exemple la distance euclidienne d'une école, ne reflète pas nécessairement la réalité du voisinage, ni le raisonnement des acheteurs types. En supposant que Jean et Janice sont représentatifs des acheteurs types dans ce secteur, force est d'admettre que la zone définie par la commission scolaire, la distance de marche et le sentiment de sécurité de leurs enfants, prévalent largement sur la distance euclidienne, qui n'exerce dans les faits aucune influence directe sur la valeur. Tout au mieux, dans un MPH, la distance euclidienne de l'école peut servir de variable proxy, c'est-à-dire une variable endogène, corrélée avec les variables omises, instaurée de manière à limiter un biais d'omission. Or, si la zone définie par la commission scolaire et la distance de marche sont des caractéristiques observables et faciles à modéliser, qu'en est-il du sentiment de sécurité de leurs enfants? Celui-ci se veut le fruit de la perception, de Jean et Janice, à l'effet que l'absence de trottoirs, l'étroitesse des rues, le nombre élevé de véhicules, et la limite de vitesse compte tenu de ces facteurs, ainsi que l'absence de sentiers piétonniers qui auraient pu compenser tous ces facteurs négatifs, compromettent la sécurité de leurs enfants. Nous avons déjà évoqué que ces facteurs sont trop nombreux pour être modélisés, en raison du problème des paramètres incidents.

Sur la base des précédentes discussions, nous posons l'emplacement comme une entité essentiellement intangible, et modifions l'équation 1, en la suivante, de manière à refléter que le prix de vente Y, s'exprime comme une fonction de l'ensemble des caractéristiques observables d'une propriété, et de la valeur intrinsèque de son emplacement :

$$Y = X\beta + E_{(v,\,z)}\,\Omega + \mu \qquad \text{(équation 2)}$$

où β est un vecteur contenant les prix hédoniques de chacune des caractéristiques observables X, Ω est un vecteur contenant les prix hédoniques de chaque emplacement E localisé aux coordonnées géographiques MTM (Mercator transverse modifiée) v et z, et μ constitue le terme d'erreur aléatoire identiquement et indépendamment distribué.

1.2.2 Les régressions géographiquement pondérées : une évolution des MPH

Les RGP constituent une évolution des MPH classiques. Elles se fondent sur la prémisse que les relations peuvent varier dans l'espace. *Ergo*, les coefficients peuvent eux aussi fluctuer selon la localisation, et donc s'avérer non stationnaires. À la sous-section 1.1.1.1, nous avons référé à ce phénomène en tant qu'hétérogénéité spatiale. Les RGP tentent de remédier à cette problématique en dégageant une série de coefficients dits locaux, à raison d'une équation de régression par emplacement.

Il est intéressant de noter que les RGP se veulent très similaires aux moindres carrés pondérés. En effet, nous rappelons, qu'en forme matricielle, le MCRL basé sur les MCO se formule de la manière suivante :

$$Y = X\beta + \mu$$

où β est un vecteur des paramètres à estimer pour chacune des variables explicatives X, et μ constitue le terme d'erreur aléatoire identiquement et indépendamment distribué.

Or, dans le cadre du MCRL, β se veut constant sur l'ensemble du territoire et peut être déterminé par l'équation matricielle suivante :

$$\beta = (X^T X)^{-1} X^T Y.$$

La différence fondamentale entre les RGP et le MCRL survient dans le calcul de l'estimateur de β. Dans sa forme matricielle, le calcul de β des RGP, au point de régression u, est impacté par la matrice de pondération géographique W(u) :

$$\beta_u = (X^T W(u) X)^{-1} X^T W(u) Y.$$

L'équation qui précède fait clairement ressortir la ressemblance entre les moindres carrés pondérés et les RGP, à la différence près que, dans le cadre de ces dernières, W est tributaire de la distance géographique entre chacune des observations et le point de régression u (Charlton et al., 2009, p. 5).

Les RGP font partie de la famille des régressions locales. Appliquées à l'immobilier, celles-ci reprennent la même forme que le MPH classique, c'est-à-dire que *Y* est posée comme une fonction de l'ensemble des caractéristiques observables d'une propriété. Toutefois, les RGP estiment autant d'équations de régression qu'il y a d'emplacements distincts à évaluer. En chaque point de régression, les observations sont pondérées selon une fonction inverse de la distance géographique les séparant de celui-ci. On peut donc poser les RGP comme suit :

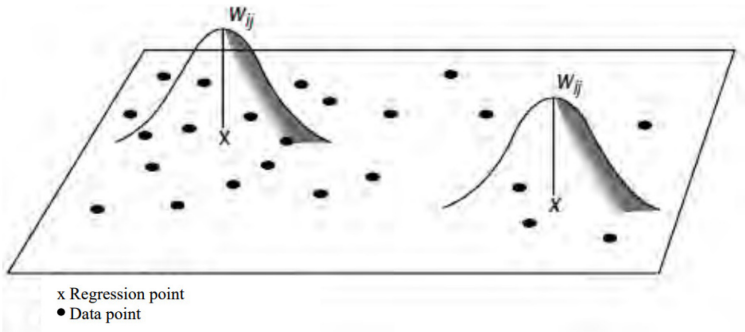
$$Y_i = \beta_{0(vi,\, zi)} + \beta_{1(vi,\, zi)} X_{i1} + \beta_{2(vi,\, zi)} X_{i2} + \ldots + \beta_{k(vi,\, zi)} X_{ik} + \mu_i \qquad \text{(équation 3)}$$

où (v_i, z_i) sont les coordonnées géographiques MTM de l'observation i et $\beta_{k(vi,\ zi)}$ est le paramètre de la variable X_k au point i notée X_{ik}.

Avant de poursuivre notre étude des RGP, il est utile de fournir certaines définitions propres à celles-ci :

« ***Bandwidth*** **»** : Il peut être défini comme étant le rayon de sélection des observations de la régression locale. Le rayon peut être fixe, c'est-à-dire que pour chaque point de régression, le modèle sélectionne et pondère toutes les observations qui sont localisées dans un rayon spécifié, souvent exprimé sous la forme d'une distance euclidienne. Ce type de rayon convient davantage aux analyses pour lesquelles la densité des observations est relativement constante sur le territoire d'analyses. La figure 1 constitue une illustration d'un rayon fixe.

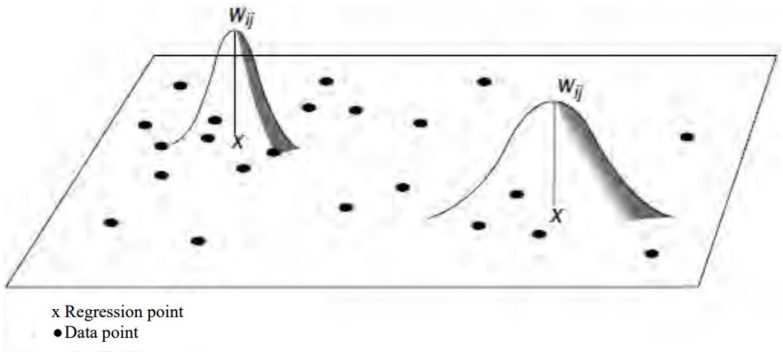
Figure 1 - Illustration d'un rayon fixe



Source : Fotheringham et al., (2002), p. 45

Le rayon peut également être adaptatif, c'est-à-dire que celui-ci s'ajuste automatiquement en chaque point de régression pour retenir un nombre équivalent d'observations. Le rayon adaptatif comporte l'avantage d'assurer de disposer d'un nombre suffisant d'observations, peu importe leur densité aux alentours d'un point de régression donné. La figure 2 illustre ce type de rayon.

Figure 2 - Illustration d'un rayon adaptatif



Source : Fotheringham et al., (2002), p. 47

Dans les RGP, le choix du rayon optimal s'exprime généralement sous la forme d'une distance euclidienne dans le cas d'un rayon fixe, et du nombre de voisins les plus proches dans le cas du rayon adaptatif. Ce choix du rayon optimal est déterminé, par essais et erreurs, de façon à optimiser un indice de performance déterminé par l'analyste. Une approche couramment utilisée porte le nom de « *cross-validation* » (CV), qui cherche à identifier le rayon h, c'est-à-dire la distance euclidienne dans le cas d'un rayon fixe ou le nombre de voisins les plus proches dans le cas d'un rayon adaptatif, qui minimise la somme du carré des résidus :

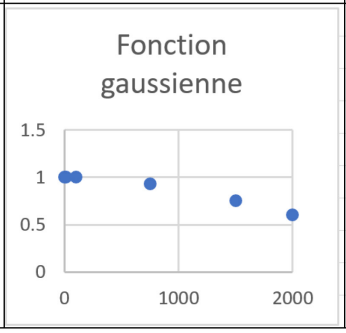
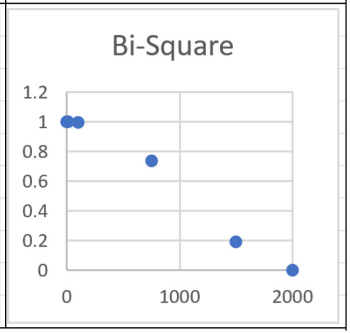
$$\mathrm{CV} = \sum_{i=1}^{n} (y_i - \hat{y}_{\neq i}(h))^2$$

où $\hat{y}_{\neq i}(h)$ est la valeur prédite de l'observation i en utilisant un rayon h et en excluant l'observation i du processus de calibration, et ce, pour chacune des observations i = 1 à i = n.

« Weighting function » : Nous référons à ce concept en tant que fonction de pondération géographique. Il s'agit essentiellement d'une fonction servant à déterminer le poids de chaque observation i, au sein d'une régression locale au point u, c'est-à-dire $w_i(u)$. La distance $d_i(u)$ de l'observation i par rapport au point de régression u, et le rayon h, sont les deux seuls déterminants de telles fonctions. Il est à noter que le

choix de la fonction de pondération géographique influence peu la performance des RGP en pratique, en autant qu'elle soit gaussienne ou quasi-gaussienne. Également, il convient de préciser que le choix du rayon se veut plus important que la fonction de pondération, en ce qui a trait à optimiser les performances des RGP (Charlton et al., 2009, p. 7). La figure 3 présente la comparaison de deux fonctions de pondération géographique : une fonction gaussienne et une fonction quasi-gaussienne de type « *bi-square* ». Dans cet exemple, le rayon h est fixe et comporte une valeur de 2 000, c'est-à-dire qu'au point de régression u, on retient toutes les observations localisées à 2 000 mètres ou moins de celui-ci. Nous soulignons une différence fondamentale entre les deux fonctions. En effet, si les observations localisées le plus près du point de régression (par exemple : 1 mètre, 10 mètres et 100 mètres) tendent à générer des pondérations similaires, nous ne pouvons en dire autant à mesure que s'accroît la distance (par exemple : 750 mètres, 1 500 mètres et 2 000 mètres). Nous remarquons, à cet effet, que la pondération de l'observation pour laquelle $d_i(u)$ = 2 000 avec la fonction « *bi-square* » est de 0, tandis qu'elle comporte une pondération significative de 0,60653066 avec la fonction gaussienne.

Figure 3 - Comparaison de deux fonctions de pondération géographique pour déterminer wi(u)

Fonctions de pondération ⇒	$w_i(\mathbf{u}) = e^{-0.5\left(d_i(\mathbf{u})/h\right)^2}$	$w_i(\mathbf{u}) = (1-(d_i(\mathbf{u})/h)^2)^2$
Distance en mètres du point de régression	Fonction Gaussienne	Bi-Square (Quasi-gaussienne)
1	0.999999875	0.9999995
10	0.9999875	0.999950001
100	0.998750781	0.99500625
750	0.932102492	0.738525391
1500	0.754839602	0.19140625
2000	0.60653066	0
Représentation des poids ⇒		

Pour conclure cette présentation des RGP, il importe de noter qu'en présence d'hétérogénéité spatiale, celles-ci produisent des estimateurs inévitablement biaisés (Fotheringham et al., 2002, p. 52). Ce biais tend à augmenter à mesure que s'accroît le rayon, du fait que les régressions incluent des observations plus lointaines du point de régression, donc davantage sujettes à l'hétérogénéité spatiale. Cependant, le fait de retenir un nombre accru d'observations réduit la variance des estimateurs. Ainsi, les RGP imposent un choix à l'analyste : celui de minimiser le biais ou encore la variance des estimateurs. Ce dilemme est connu sous le nom du compromis biais / variance dans la littérature (Fotheringham et al., 2002, p. 62 et 63).

1.2.3 La segmentation *fuzzy* dans les MPH

Nous avons évoqué, en introduction, que les propriétés immobilières sont des biens hétérogènes, voire uniques en ce qui a trait à leur emplacement. Or, ces caractéristiques rendent la segmentation *fuzzy* pour le moins attrayante d'un point de vue théorique, dans l'optique de générer une segmentation territoriale dans les MPH. À cet effet, nous avons abordé, à la sous-section 1.1.4, la théorie des ensembles *fuzzy*, qui postule qu'il existe une incertitude quant à l'appartenance à un segment donné. En évaluation immobilière, on peut facilement concevoir qu'une propriété localisée à la jonction de deux segments de voisinage puisse appartenir à la fois à ces deux segments, et ce, avec un coefficient d'appartenance associé. C'est explicitement ce que permet la segmentation *fuzzy*, qui génère de tels coefficients, à raison d'une valeur contenue entre 0 et 1, pour chaque propriété à classer, et envers tous les segments générés.

Un algorithme populaire développé pour réaliser une segmentation *fuzzy* est le *fuzzy c-means* (Dunn 1973, Bezdek 1981) qui s'inspire des nuées dynamiques, tout en permettant à chaque propriété d'appartenir à plus d'un segment. Pour ce faire, l'algorithme applique d'abord la procédure des nuées dynamiques pour générer les segments. Ensuite, il détermine des coefficients d'appartenance variant entre 0 et 1,

pour chaque observation et à l'endroit de chaque segment généré, de manière à minimiser la somme du carré des distances entre celle-ci et le centre de chaque segment.

Finalement, les coefficients d'appartenance générés pour chaque segment sont ensuite convertis en variables explicatives, qui sont elles-mêmes intégrées dans un MPH. Nous posons donc :

$$Y = X\beta + \{\sigma_1 F_1 + \sigma_2 F_2 + \ldots \sigma_k F_k\} + \mu \qquad \text{(équation 4)}$$

où β est un vecteur contenant les paramètres de chacune des caractéristiques observables X, σ_k est le paramètre relatif aux coefficients d'appartenance F_k du segment k, pour tout segment j = 1 à k, et μ constitue le terme d'erreur aléatoire identiquement et indépendamment distribué.

Voilà qui conclut le premier chapitre portant sur le cadre conceptuel de notre étude. Nous proposons maintenant de passer au second chapitre, qui traite de la démarche méthodologique suivie.

CHAPITRE 2 : MÉTHODOLOGIE

Dans le premier chapitre, nous avons abordé les principaux concepts *a priori* s'appliquant à la présente étude. Or, il importe maintenant d'adapter ceux-ci à notre contexte, et de préciser comment nous comptons les appliquer.

2.1 Mise en contexte de la présente étude

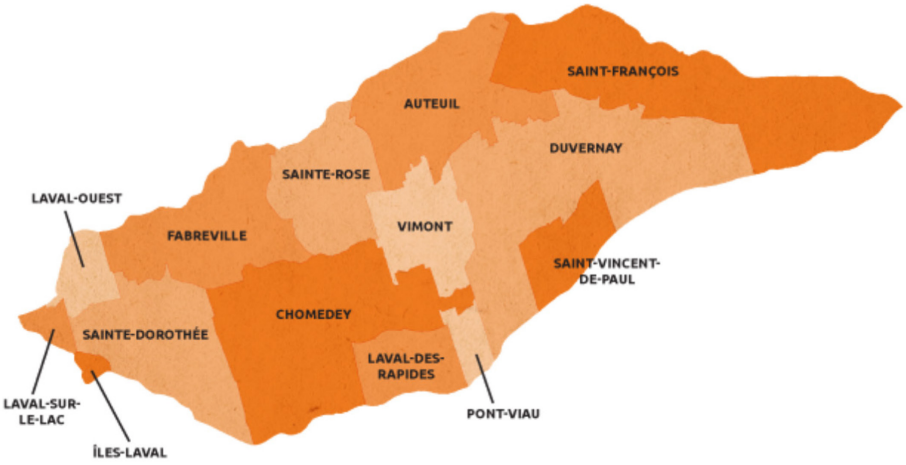
Le présent mémoire se veut une étude empirique réalisée en contexte d'évaluation municipale, au sein de la ville de Laval. Afin de bien situer le lecteur, il est de mise de procéder à une brève description de cette municipalité, ainsi qu'à une mise en contexte de la présente étude par rapport au système québécois de la taxation basée sur la valeur foncière.

2.1.1 La ville de Laval

Dans un premier temps, la ville de Laval a officiellement été fondée en 1965, année commémorant la fusion de 14 ex-municipalités en une seule grande ville unifiée. Celle-ci est la troisième ville en importance au Québec, après Montréal et Québec, en termes de population, qui se chiffre à 434 998 habitants. Elle comporte une superficie totale de 266,81 km^2 et une superficie terrestre de 246,14 km^2. La densité de sa population est donc de 1 767 habitants par km^2 terrestre.[2] La figure 4 présente les limites administratives des 14 ex-municipalités de Laval, avant la fusion municipale de 1965.

[2] https://www.mamot.gouv.qc.ca/recherche-avancee/fiche/municipalite/65005/

Figure 4 - Limites administratives des 14 ex-municipalités de Laval



Source : https://www.laval.ca/histoire-et-patrimoine/Pages/Fr/accueil.aspx

2.1.2 Le système québécois de la taxation basée sur la valeur foncière

Dans un second temps, il est pertinent de décrire sommairement le système québécois de la taxation municipale basée sur la valeur réelle. Celui-ci constitue la pierre angulaire du financement des municipalités depuis bon nombre d'années. Déjà en 1965, le rapport Bélanger faisait état des avantages d'un tel système, à savoir : son caractère local, sa simplicité d'administration et de perception pour des municipalités de toute taille et ses possibilités d'évasion fiscale restreintes. Ce même rapport identifiait néanmoins certaines lacunes, moins en lien avec la nature profonde du système que son application, à savoir notamment : l'incohérence et parfois l'absence de méthode d'évaluation et le fait que la plupart des municipalités n'évaluaient pas les propriétés à 100% de leur valeur réelle (Rapport Bélanger, 1965, p. 291 et 292). Les recommandations de ce rapport ont entraîné, dans les années qui ont suivi, d'importantes modifications au système de l'époque visant à en améliorer l'uniformité, la rigueur et l'équité. À cet effet, il importe de noter que, depuis 1979, la Loi sur la fiscalité municipale édicte que seul un évaluateur professionnel, membre de l'Ordre des évaluateurs agréés du Québec (OÉAQ), peut agir à titre d'évaluateur municipal au sein d'un organisme municipal responsable de l'évaluation (Loi sur la fiscalité

municipale, article 22). L'OÉAQ a été fondé en 1969 et « a pour mission de protéger le public en garantissant la qualité des actes professionnels posés par ses membres »[3]. L'évaluateur municipal contemporain est donc un professionnel accrédité, qui se voit assujetti à un code de déontologie et des normes de pratique professionnelle, ainsi qu'à de nombreux règlements et lois : notamment la loi sur la fiscalité municipale et le règlement sur le rôle d'évaluation foncière. L'équité se veut la valeur fondamentale du système québécois de la taxation municipale basée sur la richesse foncière. Une telle équité présume à la fois une bonne précision des évaluations, qui se doivent de refléter le mieux possible la valeur réelle de chaque immeuble à la date d'évaluation, mais aussi une excellente uniformité du processus d'évaluation qui requiert d'appliquer les mêmes conclusions et résultats d'évaluation à des groupes homogènes de propriétés situées à proximité.

2.1.3 Terminologie utilisée en évaluation immobilière

Dans un troisième temps, il s'avère pertinent de définir d'emblée certains termes propres à l'évaluation immobilière, qui seront utilisés tout au long de cet ouvrage :

Évaluation foncière : Il s'agit de la discipline professionnelle qui consiste à analyser objectivement les conditions du marché immobilier pour établir la valeur foncière de biens immobiliers[4].

Valeur foncière : Celle-ci se définit comme une opinion motivée de la valeur d'un bien immobilier énoncée à une fin particulière et à une date de référence donnée[5].

[3] https://oeaq.qc.ca/lordre/mission-protection-du-public/
[4] Manuel d'évaluation foncière du Québec, 2018
[5] Ibid.

Valeur réelle : En contexte d'évaluation municipale, on s'intéresse exclusivement à la valeur réelle qui se définit comme suit :

« La valeur réelle d'une unité d'évaluation est sa valeur d'échange sur un marché libre et ouvert à la concurrence, soit le prix le plus probable qui peut être payé lors d'une vente de gré à gré dans les conditions suivantes:

1° le vendeur et l'acheteur désirent respectivement vendre et acheter l'unité d'évaluation, mais n'y sont pas obligés; et

2° le vendeur et l'acheteur sont raisonnablement informés de l'état de l'unité d'évaluation, de l'utilisation qui peut le plus probablement en être faite et des conditions du marché immobilier. »[6]

Unités de voisinage : Celles-ci peuvent être définies comme des ensembles de propriétés connexes ou voisines, et possédant des traits communs. En contexte d'évaluation municipale, l'application des résultats est principalement gérée via les unités de voisinage. Ces dernières visent essentiellement à apparier les mêmes résultats d'évaluation à des propriétés similaires et localisées à proximité. Pour illustrer leur rôle prépondérant dans le processus d'évaluation, il convient de reprendre les mots contenus dans le Manuel d'évaluation foncière du Québec de 2018, à l'effet que : « Adéquatement réalisée et utilisée, l'unité de voisinage constitue la clé de voûte de l'équité d'un rôle d'évaluation foncière ». Cette assertion forte réaffirme non seulement le rôle crucial de la localisation, mais aussi celui de la segmentation *a priori* des unités d'évaluation pour produire des estimations à la fois transparentes, objectives, motivées et équitables, en contexte d'évaluation à une fin de taxation municipale.

[6] Loi sur la fiscalité municipale, article 43

2.1.4 Méthodes reconnues en évaluation immobilière

Dans un quatrième temps, nous poursuivons cette mise en contexte par une description des différentes méthodes utilisées en évaluation immobilière. À cet effet, il importe de noter que la doctrine fait état de trois méthodes d'évaluation : la méthode du coût, la méthode de la comparaison et la méthode du revenu. Chacune est adaptée à un contexte d'évaluation donné : notamment le but de l'évaluation, le type de propriété évaluée, la présence ou non de comparables transigés et le nombre de tels comparables.

Méthode du coût : Cette méthode d'évaluation considère une propriété comme une entité composée de deux éléments devant être évalués séparément, soit le terrain et le bâtiment :

$$V_{propriété} = V_{terrain} + (C_{constructions} - D_{constructions})$$

où $C_{constructions}$ se veut le coût de remplacement de l'ensemble des constructions (bâtiments, améliorations d'emplacement, annexes, dépendances, etc.) à la date d'évaluation et $D_{constructions}$ constitue l'ensemble des dépréciations physiques (âge et entretien), fonctionnelles (liées à l'utilité) et économiques (liées à l'environnement), affectant les constructions. La valeur du terrain $V_{terrain}$ est estimée à la date d'évaluation et ne subit aucune dépréciation. La figure 5 qui suit présente un exemple de l'application de la méthode du coût.

Figure 5 - Exemple de l'application de la méthode du coût

Adresse: [redacted]	
Ste-Rose	
Laval	
Valeur du terrain:	200 000$
Coût neuf des constructions:	275 000$
Dépréciation physique:	(20 000$)
Dépréciation fonctionnelle:	0$
Dépréciation économique:	0$
Valeur selon la méthode du coût	455 000$

Méthode de la comparaison : Cette méthode d'évaluation se décline en deux principales techniques : la technique des prix de vente ajustés et l'approche hédonique.

Technique des prix de vente ajustés : Celle approche préconise de comparer le sujet, c'est-à-dire la propriété à évaluer, à des propriétés similaires ayant fait l'objet d'une transaction : les comparables. La sélection des comparables est généralement déterminée par l'analyste sur la base des ressemblances (caractéristiques physiques, proximité, etc.) par rapport au sujet, et la date de la transaction, qui se doit d'être le plus près possible de la date d'évaluation. En pratique, on retient généralement de trois à dix comparables, qui se doivent dès lors d'être très similaires au sujet, en raison de leur faible nombre. Une fois les comparables sélectionnés, une série d'ajustements sont appliqués aux prix de vente, afin de rendre ces derniers plus représentatifs de la valeur du sujet, que ce soit pour actualiser les prix à la date d'évaluation, prendre en compte un contexte de vente particulier tel une succession ou un divorce, ou encore ajuster les prix en fonction des différences constatées dans les caractéristiques physiques et de localisation. Les prix de vente ajustés obtenus constituent des indicateurs de valeur du sujet, à partir desquels l'analyste établit sa conclusion de valeur. À cette fin, il peut choisir de retenir la moyenne, la médiane ou le mode des indicateurs de valeur, ou encore d'ajuster leur pondération dans ses calculs. Peu importe son choix, il doit être en mesure de le justifier. La figure 6 présente un exemple de l'évaluation d'une

propriété par la technique des prix de vente ajustés. On compare un sujet à trois comparables, dont les prix de vente respectifs de 500 000$, 453 000$ et 539 000$ sont ajustés pour tenir compte des divergences dans les caractéristiques. Une fois ajustés, les prix de vente produisent des indicateurs de valeur respectifs de 477 200$, 454 400$ et 481 500$. Dans le cas présent, les ajustements sont quantifiés sur la base des estimateurs fournis par l'approche hédonique. Une approche alternative consiste à estimer ces ajustements sur la base du coût déprécié des composantes.

Figure 6 - Exemple démontrant l'application de la technique des prix de vente ajustés

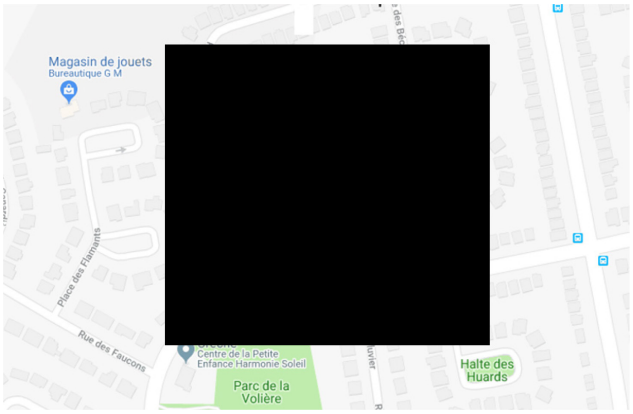
	SUJET		1		2		3	
								
IDF-BAT	[redacted]		[redacted]		[redacted]		[redacted]	
Adresse								
Matricule								
Distance			0,0243 KM		0,0355 KM		0,0487 KM	
No Enregistrement			[redacted]		[redacted]		[redacted]	
Date de vente			2017-11-30		2016-03-03		2018-05-09	
Prix de vente			500000 $		453000 $		539000 $	
Prix de vente rajusté			500000 $		453000 $		539000 $	
Facteur d'actualisation			1		1,029550		0,957190	
Prix de vente rajusté actualisé			500000 $		466400 $		515900 $	
Rajustements des comparables			-22843 $		-12010 $		-34416 $	
Indicateur de valeur			477200$		454400$		481500$	
Alias(Coefficient$)	**Val. alias**	**Résultat**	**Val. alias**	**Rajustement**	**Val. alias**	**Rajustement**	**Val. alias**	**Rajustement**
AIRE_HAB_UNI_LN(0,3513999780)	4,98	1,75	4,99	-0,006249	5,06	-0,029523	5,05	-0,026613
IF_1_SDB(-0,0202963420)	0	0,00	0	0,000000	0	0,000000	0	0,000000
IF_ASP_CENTRAL(0,0092074930)	1	0,01	1	0,000000	1	0,000000	0	0,009207
IF_FOYER(0,0124292570)	0	0,00	0	0,000000	0	0,000000	0	0,000000
IF_PISCINE_HT(0,0088208020)	0	0,00	0	0,000000	0	0,000000	0	0,000000
POURC_BRIQUE_PIERRE(0,0002353450)	20	0,00	18	0,000471	23	-0,000706	14	0,001412
POURC_VINYLE(-0,0001563430)	50	-0,01	42	-0,001251	51	0,000156	53	0,000469
age_app_2017(-0,0091693540)	6	-0,06	6	0,000000	6	0,000000	5	-0,009169
aire_gar_att_int(0,0018320690)	22,40	0,04	28,90	-0,011908	22,50	-0,000183	24	-0,002931
clim_central(0,0323675780)	0	0,00	1	-0,032368	1	-0,032368	1	-0,032368
clim_mural_bin(0,0135919740)	1	0,01	0	0,013592	0	0,013592	0	0,013592
constante(10,28831359)	1	10,29	1	0,000000	1	0,000000	1	0,000000
pisc_creus_bin(0,0393105760)	0	0,00	0	0,000000	0	0,000000	1	-0,039311
planch_sup_pourc(0,0002303570)	100	0,02	100	0,000000	100	0,000000	100	0,000000
ssol_fini_aire(0,0005764280)	65,70	0,04	55,50	0,005880	0	0,037871	65,90	-0,000115
CLASSE_POINT_TOTAL(0,0021280060)	106	0,23	113	-0,014896	113	-0,014896	102	0,008512
SUP_TERRAIN_LN(0,1199747610)	5,92	0,71	5,92	-0,000032	5,92	-0,000032	5,85	0,008277
uv_4570_bin(0,01426336)	0	0,00	0	0,000000	0	0,000000	0	0,000000

Source : Système de la comparaison directe de la ville de Laval

Par ailleurs, il est d'usage d'accompagner le tableau qui précède d'une carte localisant à la fois le sujet et les comparables retenus. La figure 7 fournit un exemple

d'une telle carte. On y aperçoit le sujet (S) et les trois comparables retenus (numérotés de 1 à 3).

Figure 7 - Exemple de carte localisant le sujet et les comparables



Source : Google Maps

La principale force de la technique des prix de vente ajustés consiste à préconiser des propriétés généralement très similaires au sujet et localisées à proximité. Son tendon d'Achille est, en quelque sorte, la contrepartie de sa principale force, à savoir que les très bons comparables, pour une propriété donnée, sont généralement limités en nombre. Or, en analyse statistique, on sait que le faible nombre d'observations a pour effet d'accroître la variance des estimations. Qui plus est, malgré les ressemblances dans les caractéristiques physiques et de l'emplacement, rien n'indique que le prix payé, pour une maison donnée, est forcément représentatif du prix que paieraient plusieurs acheteurs potentiels pour cette même propriété. À cet effet, il importe de mentionner qu'une transaction immobilière donnée implique inévitablement des aléas non mesurables, tels que les habiletés de négociation des parties, des facteurs émotionnels divers, un contexte situationnel particulier des parties, une connaissance variable du marché immobilier et le hasard lui-même. Il s'ensuit que le comparable le plus similaire, en termes de caractéristiques physiques, n'est pas forcément le plus représentatif de la valeur du sujet.

Approche hédonique : Cette technique se fonde sur les MCO et préconise les échantillons plus larges afin d'accroître la robustesse statistique des estimations. Elle vise essentiellement à décomposer les prix de vente des comparables en une série de valeurs contributives pour les divers attributs les constituant. Une fois ces valeurs contributives dégagées, l'analyste peut les appliquer aux caractéristiques d'une propriété à évaluer. La figure 8 présente un exemple, non lié à la présente étude, de l'application de l'approche hédonique. La forme fonctionnelle illustrée est de type semi-log, c'est-à-dire que la variable dépendante se veut le logarithme népérien du prix de vente. Nous y remarquons que la quantité de chaque caractéristique est multipliée par son estimateur associé. En forme semi-log, on procède ensuite à la somme de tous les résultats de ces multiplications, à laquelle on additionne la valeur de la constante, puis on applique une fonction inverse sur ce total. Le résultat de chaque estimation est ainsi converti en dollars.

Figure 8 - Exemple de l'application de l'approche hédonique sous une forme semi-log

Total	460100 $
IDF	
Bâtiment	
Adresse	
Matricule	
Parc immobilier	Unifamilial - Rôle 2019

Calcul des coefficients

Alias	Paramètre	Coefficient	Résultat
age_app_2017	6	-0.009169	-0.0550
aire_gar_att_int	22.4	0.001832	0.0410
AIRE_HAB_UNI_LN	4.976	0.3514	1.7490
CLASSE_POINT_TOTAL	106	0.002128	0.2260
clim_mural_bin	1	0.013592	0.0140
constante	1	10.288314	10.2880
IF_ASP_CENTRAL	1	0.009207	0.0090
planch_sup_pourc	100	0.000230	0.0230
POURC_BRIQUE_PIERRE	20	0.000235	0.0050
POURC_VINYLE	50	-0.000156	-0.0080
ssol_fini_aire	65.7	0.000576	0.0380
SUP_TERRAIN_LN	5.919	0.119975	0.71

Forme Semi-log: EXP(6 * -0.009169 + 22.4 * 0.001832 + 4.976 * 0.3514 + 106 * 0.002128 + 1 * 0.013592 + 1 * 10.288314 + 1 * 0.009207 + 100 * 0.000230 + 20 * 0.000235 + 50 * -0.000156 + 65.7 * 0.000576 + 5.919 * 0.119975)

Source : Système de la comparaison directe de la ville de Laval

Méthode du revenu: Cette méthode d'évaluation concerne les immeubles générateurs de revenus. Celle-ci assimile la valeur d'une propriété immobilière en fonction des flux anticipés de ses revenus nets. La valeur actuelle de ces derniers constitue l'indice de valeur dégagé par cette méthode. Il est à noter que cette dernière n'est pas applicable dans l'évaluation d'immeubles résidentiels de type unifamilial indivis, tels que ceux faisant l'objet de la présente étude.

2.1.5 Adoption de la méthode de la comparaison en évaluation municipale

Pour conclure cette mise en contexte, il importe de noter, qu'en pratique, l'évaluateur est souvent confronté à choisir la meilleure méthode d'évaluation pour un immeuble donné. Pour reprendre les mots du professeur Mario Fortin, de l'Université de Sherbrooke, celui qui possède une seule montre connaît toujours l'heure, tandis que celui qui en possède plusieurs n'en est jamais vraiment certain. En évaluation immobilière, en présence d'un nombre adéquat de comparables, il est d'usage de préconiser les approches qui ciblent directement le comportement des vendeurs et acheteurs sur le marché immobilier. De telles approches sont dites directes. Or, en évaluation immobilière, la méthode de la comparaison est la seule qui soit unanimement reconnue comme une approche directe. Il s'ensuit qu'en présence d'un nombre suffisant de comparables, celle-ci constitue la méthode par excellence en évaluation immobilière.

Par ailleurs, dans le contexte de la plus récente modernisation réglementaire et normative de l'évaluation foncière, pilotée par Ministère des Affaires municipales et Occupation du territoire, l'Ordre des évaluateurs agréés du Québec exige désormais des évaluateurs municipaux, via la norme de pratique 20.1 adoptée le 31 janvier 2013, de sélectionner et justifier « quelle(s) méthode(s) d'évaluation est (sont) la (les) plus pertinente(s) » pour une unité d'évaluation donnée. C'est principalement l'instauration de cette norme de pratique qui a contribué à l'émergence de l'utilisation de la méthode

de la comparaison et de l'approche hédonique en contexte d'évaluation municipale au Québec, traditionnellement fortement ancrée dans la méthode du coût.

2.2 Définition des propriétés faisant l'objet de la présente étude

Cette section définit, de manière concise, les propriétés faisant l'objet de la présente étude.

Tout d'abord, il convient de noter que les jeux de données obtenus, pour la ville de Laval, se scindent en deux catégories : celle des propriétés transigées qui, dans le cadre du présent mémoire, sont désignées en tant que ventes, et les propriétés à évaluer qui sont référées en tant qu'inventaire. Du point de vue de l'inférence statistique, l'inventaire réfère à la population, c'est-à-dire l'ensemble des propriétés à évaluer, tandis que les ventes constituent l'échantillon, issu de la population et présumé aléatoire et indépendant, à partir duquel on peut générer des résultats d'évaluation, et les généraliser à la population cible. Nous croyons qu'il est pertinent, en pratique, de distinguer l'inventaire des ventes pour deux raisons principales :

1. **Exhaustivité :** Pour être en mesure d'évaluer toutes les propriétés de l'inventaire, et non uniquement celles comportant des caractéristiques représentées au sein des ventes. Par exemple, une analyse descriptive de l'inventaire peut permettre d'identifier des unités de voisinage ne comportant aucune vente, ce qui peut inciter à revoir la stratégie de segmentation territoriale;

2. **Juger du niveau d'extrapolation :** L'extrapolation consiste à produire des estimations sur des propriétés dont une ou plusieurs caractéristiques ne sont pas représentées parmi les ventes. En guise d'exemple, une analyse descriptive de l'inventaire peut permettre de constater la présence d'une maison à évaluer dont l'aire habitable excède de près du double celle de la plus grande maison

transigée. L'analyste peut ainsi choisir d'exclure de telles propriétés de l'inventaire, ou encore de marquer celles-ci afin d'y porter une attention particulière au moment de conclure ses analyses.

Or, nous avons évoqué précédemment que la conjugaison de propriétés trop disparates au sein d'un même MPH peut engendrer diverses problématiques d'ordre économétrique. Pour cette raison, la présente étude cible exclusivement les résidences de type unifamilial indivis, attachées et détachées, de qualité et de complexité relativement standard, c'est-à-dire comportant un pointage de classe se situant entre 64 et 138, selon la procédure décrite dans le Manuel d'évaluation foncière du Québec de 2018. Seules les propriétés dont le code d'utilisation est de 1000, c'est-à-dire les logements, et desservies par les services d'égouts et d'aqueduc de la municipalité sont retenues. Nous excluons les propriétés exemptes de ces services, en raison de leur faible nombre, de leur très forte autocorrélation spatiale, de leur forte multicolinéarité avec d'autres variables explicatives, et par souci d'une meilleure homogénéité dans les modèles. Les copropriétés divises, les propriétés riveraines et les maisons mobiles sont également exclues de la présente analyse, puisqu'elles constituent des marchés distincts de celui des résidences unifamiliales étudiées dans le cadre du présent mémoire.

En ce qui concerne les ventes, nous retenons l'ensemble des transactions *bona fide* survenues entre le 1er janvier 2016 et le 1er juillet 2018. La première date commémore l'entrée en vigueur de la modernisation réglementaire et normative, abordée à la sous-section 2.1.5, et la migration des données de l'ancien système au nouveau, au service d'évaluation de la ville de Laval. Cette migration des données a entraîné une rupture au niveau de certains renseignements, par exemple le calcul des classes, ce qui nous contraint à restreindre la présente étude aux observations ultérieures à celle-ci. Quant à la date du 1er juillet 2018, celle-ci est retenue pour une raison pratique, reliée à l'échéancier de la conception du rôle d'évaluation, qui doit être déposé au plus tard au mois d'octobre 2018.

Par ailleurs, seules les ventes *bona fide* admissibles, et pour lesquelles le contexte de vente est jugé normal, sont retenues aux fins de la présente étude. C'est donc dire que les ventes en contexte de succession, de préavis d'exercice hypothécaire, de séparation ou divorce, de transfert du vendeur ou toute autre situation ne s'assimilant pas aux conditions normales de vente sont exclues. Il en est de même pour les ventes se rapportant à un immeuble transigé sans la garantie légale de qualité, ayant nécessité d'importants travaux de transformation, ayant servi à la plantation de cannabis ou dans laquelle il y a eu une mort violente. Finalement, aucun filtre n'est appliqué sur les prix de vente : l'échantillonnage endogène est ainsi écarté afin de ne pas biaiser les estimateurs (Wooldridge, 2013, p. 325).

2.3 Stratégie de l'évaluation de la performance

En introduction, nous avons défini le concept de surapprentissage, et avons évoqué que, pour être utile, un modèle doit générer une bonne performance, non seulement sur les observations utilisées pour le calibrer, mais surtout sur des données qui n'ont pas encore été observées.

Dans cette optique, nous évaluons la performance prédictive des modèles strictement sur la base d'observations non utilisées pour les calibrer. Nous préconisons une technique de validation croisée sur les indicateurs de performance comportant 100 itérations. Une telle manière de procéder assure que la performance mesurée n'est pas le fruit du surapprentissage, ou encore de la composition d'un échantillon donné.

Dans cette étude, chacune des 100 itérations de la technique de validation croisée sur les indicateurs de performance consiste à isoler aléatoirement 10% des observations à titre d'échantillon de test, à calibrer les modèles exclusivement sur la base des observations non isolées correspondant à 90% du jeu de données original, et finalement à mesurer la performance sur les observations isolées contenues dans l'échantillon de test. Chaque échantillon de test comporte 460 observations, tandis que

chaque échantillon d'apprentissage en compte 4 132. Pour chaque indicateur de performance calculé sur la base de cette technique, nous retenons la moyenne obtenue sur les 100 itérations.

Par ailleurs, dans le but d'évaluer la généralisation des modèles, c'est-à-dire leur propension au surapprentissage, nous calibrons également ceux-ci en utilisant l'ensemble des ventes, et calculons les indicateurs de performance sur la base de ces 4 592 ventes, sans égard au fait qu'elles servent à la fois à calibrer les modèles et à mesurer la performance. Pour chaque modèle, l'écart entre les indicateurs de performance obtenus à partir de l'ensemble des ventes, et ceux générés par la technique de validation croisée, permet d'évaluer leur généralisation.

2.4 Indicateurs de performance sélectionnés

Nous présentons maintenant les indicateurs de performance sélectionnés dans le cadre de la présente étude. Les indicateurs I1 à I7 sont calculés à la fois sur la base de la technique de validation croisée et sur l'ensemble des ventes, tel qu'expliqué à la section 2.3. Quant aux indicateurs I8 à I10, qui se veulent des indicateurs statistiques plus classiques, nous les calculons exclusivement sur l'ensemble des 4 592 observations.

Tous les indicateurs de performance sont calculés sur la base des ventes i = 1 à n_v où n_v est le nombre de ventes à partir duquel on calcule l'indice. Nous rappelons que dans le cadre de la technique de validation croisée, $n_v = 460$ pour chaque itération, tandis que $n_v = 4\ 592$ lorsque les modèles sont calibrés sur l'ensemble des ventes. Pour chaque indicateur de performance applicable, Y_i désigne le prix de vente réel de la vente i, $\hat{Y}_i$ se veut l'estimation du modèle pour la vente i, et $\bar{Y}$ est le prix moyen de l'ensemble des ventes retenues pour calculer un indicateur de performance donné. Finalement, k désigne le nombre de variables explicatives pour un modèle donné, excluant la constante.

I1. Proportion médiane des rapports estimation / prix de vente : Le règlement sur la proportion médiane du rôle d'évaluation foncière prescrit l'utilisation de cet indice de performance en contexte d'évaluation municipale. Plus spécifiquement, il prévoit de calculer, pour chaque propriété vendue admissible, le quotient de la valeur inscrite au rôle de l'unité d'évaluation par le prix de vente. On retient ensuite la médiane des rapports calculés, c'est-à-dire :

$$I1 = \text{médiane des rapports } (r_1, r_2, \ldots, r_i)$$

où r_1, r_2, …, r_i sont les rapports estimation / prix de vente de chacune des ventes.

I2. Écart type relatif à la médiane : L'annexe III du règlement sur le rôle d'évaluation foncière décrit la manière de calculer cet indicateur. Il s'agit d'une mesure de la dispersion des ratios estimation / prix de vente, donc une indication de la précision des estimations. On l'obtient en appliquant la formule suivante :

$$I2 = \left(\sqrt{\frac{\sum_{i=1}^{i=n_v} (r_i - I1)^2}{n_v - 1}} \right) \div I1.$$

I3. Taux d'erreur moyen en valeur absolue : Cet indice est connu en tant que MAPE ou encore « *mean absolute percentage error* ». Il permet de comparer l'ampleur des erreurs aux valeurs observées, abstraction faite de l'unité de mesure des observations. Dans le présent contexte, cet indicateur se calcule de la manière suivante :

$$I3 = \frac{\sum_{i=1}^{i=n_v} (|Y_i - \hat{Y}_i| / Y_i)}{n_v}.$$

I4. Erreur moyenne en dollars en valeur absolue : Cet indicateur, connu en tant que MAD ou encore « *mean absolute deviation* », mesure l'ampleur de l'erreur en fonction de l'unité étudiée. Dans le contexte de l'évaluation municipale, cet indicateur fournit une erreur d'estimation en dollars et se calcule ainsi :

$$\mathrm{I4} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{i}=n_v}(\left|\mathrm{Y_i} - \hat{\mathrm{Y}}_\mathrm{i}\right|)}{n_v}.$$

I5. Proportion des taux d'erreur en valeur absolue en deçà de 10% : Cet indicateur exprime la proportion des ventes dont le taux d'erreur en valeur absolue est inférieur au seuil de 10%.

$$\mathrm{I5} = \frac{\mathrm{n_{te<10\%}}}{n_v}$$

où $\mathrm{n_{te<10\%}}$ est le nombre de ventes estimées avec un taux d'erreur en deçà de 10%.

I6. Moyenne des taux d'erreur : Connu en tant que MPE ou « *mean percentage error* », cet indicateur évalue dans quelle mesure un modèle tend à surévaluer ou sous-évaluer systématiquement les prix de vente. Les résidus positifs et négatifs génèrent des taux d'erreur positifs et négatifs qui s'annulent lorsqu'ils sont additionnés. Un modèle peu ou pas biaisé comportera par conséquent un MPE très près de 0, tandis qu'un modèle biaisé se caractérisera par un MPE s'éloignant de 0.

$$\mathrm{I6} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{i}=n_v}((\mathrm{Y_i} - \hat{\mathrm{Y}}_\mathrm{i})/\mathrm{Y_i})}{n_v}.$$

I7. Index de Moran : Cet indicateur est le plus couramment utilisé pour tester l'autocorrélation spatiale dans les résidus. À l'instar du coefficient de corrélation, l'index de Moran produit une valeur entre -1 et 1. Une valeur de -1 signifie une parfaite

autocorrélation spatiale négative, c'est-à-dire que les résidus sont entourés par des valeurs diamétralement opposées. Une valeur près de 0, ou plus précisément de $\frac{-1}{(n_v - 1)}$, indique l'indépendance spatiale. Finalement, une valeur de 1 suggère une parfaite autocorrélation spatiale positive, en quel cas les résidus sont entourés par d'autres de de même signe et amplitude. En contexte d'inférence, l'hypothèse nulle se veut l'indépendance spatiale.

Suivant Anselin (1988, p. 101 et 102), le I de Moran calculé dans la présente étude est représenté par l'équation matricielle suivante :

$$\mathrm{I7} = \left[\frac{n_v}{\mathrm{S}}\right] \times \left[\frac{\mathrm{e'We}}{\mathrm{e'e}}\right]$$

où S est la somme de toutes les pondérations contenues dans la matrice de pondérations géographiques W et e est le vecteur contenant les résidus. Par ailleurs, dans le cadre du présent mémoire, la matrice des pondérations géographiques W, servant au calcul du I de Moran, est déterminée par la fonction suivante :

$$W_{ij} = \left[\frac{1}{d_{ij} + 1}\right]$$

où W_{ij} est la pondération géographique réciproque des ventes localisées aux emplacements i et j, et d_{ij} est la distance euclidienne séparant ces deux ventes.

Par ailleurs, il est intéressant de noter, dans la formule I7, comme le fait remarquer Anselin (1988, p. 102), que lorsque la somme des pondérations pour chacune des lignes de la matrice W équivaut à 1, on obtient alors $n_v = S$, faisant en sorte qu'on peut simplifier l'équation I7, pour obtenir :

$$\text{I de Moran}_{\text{si chaque ligne de W=1}} = \left[\frac{e'We}{e'e}\right].$$

Or, l'équation qui précède est mathématiquement équivalente à une régression par les moindres carrés ordinaires de We sur e (Anselin, 1988, p. 102). La pente d'une telle régression correspond donc également à la statistique I de Moran.

Par ailleurs, en ce qui concerne les indicateurs de performance I8 à I10, ceux-ci sont calculés exclusivement sur la base de l'ensemble des 4 592 observations.

I8. R^2: Cet indicateur populaire exprime la proportion de la variation empirique de la variable dépendante qui est expliquée par le modèle.

$$\text{I8} = \frac{\sum_{\text{i}=1}^{\text{i}=n_v}(\hat{\text{Y}}_\text{i} - \bar{\text{Y}})^2}{\sum_{\text{i}=1}^{\text{i}=n_v}(Y_\text{i} - \bar{\text{Y}})^2}.$$

I9. R^2 ajusté : L'ajout d'une variable explicative peut accroître le R^2, sans pour autant améliorer la qualité d'ajustement du modèle. Autrement dit, d'un point de vue statistique, la diminution de la somme du carré des résidus peut s'avérer insuffisante pour pallier la perte d'un degré de liberté occasionnée par l'ajout d'une variable explicative dans un modèle. Le R^2 ajusté compense cette lacune en ajustant le R^2 par un facteur, qui tient compte à la fois du nombre de variables explicatives et du nombre d'observations dans le modèle.

$$\text{I9} = 1 - \left(\frac{n_v - 1}{n_v - 1 - \text{k}}\right)\frac{\sum_{\text{i}=1}^{\text{i}=n_v}\left(\text{Y}_\text{i} - \hat{\text{Y}}_\text{i}\right)^2}{\sum_{\text{i}=1}^{\text{i}=n_v}\left(\text{Y}_\text{i} - \bar{\text{Y}}\right)^2}.$$

I10. Erreur-type de la régression : Il s'agit de l'estimateur de l'erreur-type du terme d'erreur de la régression. On l'obtient par le calcul suivant :

$$\mathrm{I10} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{i}=n_v} (\mathrm{Y_i} - \hat{\mathrm{Y}}_\mathrm{i})^2}{n_v - 1 - \mathrm{k}}.$$

2.5 Spécificités des modèles testés

Chacun des cinq modèles à tester comporte la même spécification. La seule distinction entre ceux-ci se rapporte à la manière dont la localisation est prise en compte. La forme fonctionnelle retenue est de type semi-log, c'est-à-dire que la variable dépendante est le logarithme népérien du prix de vente, tandis que certaines variables explicatives conservent leur forme originale, et d'autre subissent une transformation mathématique pour demeurer linéaire dans les paramètres. Une telle forme fonctionnelle génère des modèles multiplicatifs, en comparaison aux modèles linéaires additifs traditionnels.

Cette section présente donc, tour à tour, les modèles testés dans le cadre de ce mémoire. Pour chacun, la variable dépendante, Y, se veut le logarithme népérien du prix de vente de la propriété. La figure 9 présente la spécification commune aux cinq modèles testés.

Figure 9 - Spécification commune aux cinq modèles testés

Nom de la variable	Description de la variable
1. AGE_APP_2018	Âge apparent de la propriété en 2018 (c'est-à-dire l'âge originel de la propriété modifié ou non selon l'entretien et les rénovations)
2. AIRE_GAR_ATT_INT	Aire du garage attaché ou intégré en mètres carrés
3. AIRE_GAR_SSOL	Aire du garage au sous-sol en mètres carrés
4. AIRE_HAB_UNI	Aire habitable de la propriété en mètres carrés
5. AIRE_HAB_UNI_LN	Logarithme népérien de l'aire habitable de la propriété en mètres carrés
6. AIRE_SSOL_FINIE	Aire aménagée du sous-sol en mètres carrés
7. IF_1_SDB	Variable indicatrice: La propriété comporte-t-elle une seule salle de bain complète?
8. IF_ASP_CENTRAL	Variable indicatrice: La propriété comporte-t-elle un aspirateur central?
9. IF_BUNGALOW	Variable indicatrice: S'agit-il d'une propriété de type bungalow?
10. IF_CLIM_CENTRAL	Variable indicatrice: La propriété comporte-t-elle un climatiseur central?
11. IF_CLIM_MURAL	Variable indicatrice: La propriété comporte-t-elle un climatiseur mural?
12. IF_EN_RANGEE_1	Variable indicatrice: S'agit-il d'une propriété de type en rangée extérieure?
13. IF_EN_RANGEE_2	Variable indicatrice: S'agit-il d'une propriété de type en rangée intérieure?
14. IF_FOND_PAS_SSOL	Variable indicatrice: Le type de fondation dominant n'est pas un sous-sol
15. IF_FOYER	Variable indicatrice: La propriété comporte-t-elle un foyer?
16. IF_INON_0_20	Variable indicatrice: La propriété est-elle située en zone inondable 0 à 20 ans?
17. IF_INON_20_100	Variable indicatrice: La propriété est-elle située en zone inondable 20 à 100 ans?
18. IF_JUMELE	Variable indicatrice: S'agit-il d'une propriété de type jumelé?
19. IF_MANSARDE	Variable indicatrice: S'agit-il d'une propriété de type à étage mansardé?
20. IF_NIV_DECALES	Variable indicatrice: S'agit-il d'une propriété de type à niveaux décalés?
21. IF_PISCINE_EXC	Variable indicatrice: La propriété comporte-t-elle une piscine creusée?
22. IF_PISCINE_HT	Variable indicatrice: La propriété comporte-t-elle une piscine hors-terre?
23. IF_VENTE_2016	Variable indicatrice: La transaction a-t-elle été conclue en 2016?
24. IF_VENTE_2018	Variable indicatrice: La transaction a-t-elle été conclue en 2018?
25. LN_PRIX_RAJUSTE	Variable dépendante: Logarithme népérien du prix de vente
26. POINTAGE_CLASSE	Pointage de qualité et de complexité de la propriété selon la procédure du MÉFQ
27. POURC_BRIQUE_PIERRE	Pourcentage du parement des murs extérieurs en brique ou en pierre
28. POURC_PLANCHER_SUP	Pourcentage des revêtements de plancher qualifiés de supérieurs (bois franc, céramique, ardoise, marbre, etc.)
29. POURC_VINYLE	Pourcentage du parement des murs extérieurs en vinyle
30. SUP_TERRAIN_LN	Logarithme népérien de la superficie du terrain de la propriété en mètres carrés

2.5.1 MPH sans découpage territorial (NAIF)

Ce modèle est constitué exclusivement des variables que nous venons tout juste de présenter, à la figure 9. Nous omettons volontairement la dimension de la

localisation, d'une part dans le but d'évaluer la performance des segmentations territoriales testées, et d'autre part pour analyser les estimateurs en absence d'une segmentation territoriale adéquate.

2.5.2 MPH avec découpage *a priori* par des évaluateurs professionnels (EXPERT)

Nous avons précédemment souligné, qu'en théorie, en raison de leur unicité, chaque emplacement peut commander un prix hédonique spécifique. Or, nous avons également évoqué, qu'en raison du nombre limité d'observations et du problème des paramètres incidents, il est impossible en pratique d'estimer les prix hédoniques de chaque emplacement dans les MPH.

Nous rappelons qu'en contexte d'évaluation municipale, la segmentation territoriale est réalisée via le découpage d'unités de voisinage. Or, il arrive souvent, en pratique, que le nombre de ventes disponibles pour une unité de voisinage donnée soit nul ou insuffisant pour estimer les paramètres de manière fiable. Pour remédier à cette problématique, l'analyste envisage généralement deux solutions : la première est d'extrapoler les résultats provenant d'autres unités de voisinage jugées similaires, et la seconde est d'adapter le niveau d'agrégation de manière à obtenir un nombre suffisant d'observations. Dans le cadre de cette étude, nous retenons cette dernière approche, puisqu'un tel niveau d'agrégation à plus haut niveau existe déjà à la ville de Laval. À cet effet, nous distinguons deux découpages distincts, tous deux réalisés par des évaluateurs professionnels, à savoir : les unités de voisinage qui reflètent un découpage très fin du territoire, et les sous bassins d'analyses qui regroupent plusieurs unités de voisinages connexes ou voisines.

En adoptant, dans le modèle EXPERT, une segmentation *a priori* réalisée par des évaluateurs professionnels, via le découpage d'unités de voisinage et de sous bassins d'analyses, nous pouvons adapter l'équation 2, présentée à la sous-section 1.2.1, en la suivante :

$$Y = X\beta + UV\delta + SB\Phi + \mu \qquad \text{(équation 5)}$$

où β est un vecteur contenant les prix hédoniques de chacune des caractéristiques observables X, δ est un vecteur contenant les prix hédoniques des unités de voisinage UV, et Φ est un vecteur contenant les prix hédoniques des sous bassins d'analyses SB.

Bien qu'il soit généralement recommandé, dans les modèles linéaires multivariés, de disposer d'au moins 10 observations pour chacune des modalités d'une variable catégorique, il appert, dans le cas présent, qu'un tel critère restreindrait l'analyse à un nombre très limité d'unités de voisinage. À cet effet, nous notons que seulement 129 unités de voisinage parmi les 837 présentes dans l'inventaire, c'est-à-dire à peine 15,41%, comprennent un tel nombre de ventes. Qui plus est, nos tests empiriques ont démontré qu'un tel seuil n'est pas optimal, à tout le moins dans la présente étude. Nous constatons plutôt, après quelques tests, que la performance prédictive du modèle EXPERT se trouve optimisée en créant des variables binaires pour chacune des unités de voisinage comportant cinq ventes ou plus. Par conséquent, dans la présente étude, un total de 274 variables binaires sont créées pour les unités de voisinage. Quant aux 563 unités de voisinage n'atteignant pas ce seuil, nous retenons le sous bassin d'analyses en tant que variable de localisation. Un total de 31 variables binaires identifiant les sous bassin d'analyses sont donc également incluses au sein de ce modèle.

2.5.3 Les régressions géographiquement pondérées (RGP)

Nous avons abordé les RGP à la sous-section 1.2.2. Or, il convient maintenant de détailler comment nous calibrons ce modèle dans notre étude.

Choix du type de rayon : Dans un premier temps, nous préconisons un rayon adaptatif. Le choix d'un tel rayon est justifié par le fait que la densité des propriétés est

relativement variable sur le territoire de Laval. C'est donc dire que, pour chaque propriété à évaluer, nous retenons dans les RGP un rayon h qui, bien qu'exprimé sous forme de distance euclidienne, s'ajuste automatiquement pour retenir un nombre équivalent d'observations, c'est-à-dire les o observations les plus près, en chaque point de régression.

Choix de la fonction de pondération géographique : Dans un second temps, nous choisissons une fonction de pondération géographique de type « *bi-square* », représentée par la formule suivante au point de régression u :

$$w_i(u) = \left(1 - \left(\frac{d_i(u)}{h}\right)^2\right)^2$$

où $d_i(u)$ est la distance euclidienne entre l'observation i et le point de régression u, et h est la distance euclidienne entre le point de régression u et l'observation la plus éloignée retenue parmi les o observations les plus près.

Choix du rayon optimal : Finalement, dans la présente étude, nous sélectionnons le rayon adaptatif optimal sur la base de la procédure CV, qui consiste à rechercher le nombre o minimisant la somme du carré des résidus pour l'ensemble des ventes utilisées. Ce nombre optimal est noté $o_{optimal}$. Afin d'évaluer le plus justement possible la performance prédictive des RGP, nous appliquons cette procédure CV conformément à la technique de validation croisée discutée précédemment. C'est donc dire que, pour chacune des 100 itérations, la procédure CV est réalisée strictement sur la base de l'échantillon d'apprentissage. Pour une itération donnée, $o_{optimal}$ est ensuite appliqué au sein de l'échantillon de test, pour estimer les régressions et générer les indices de performance.

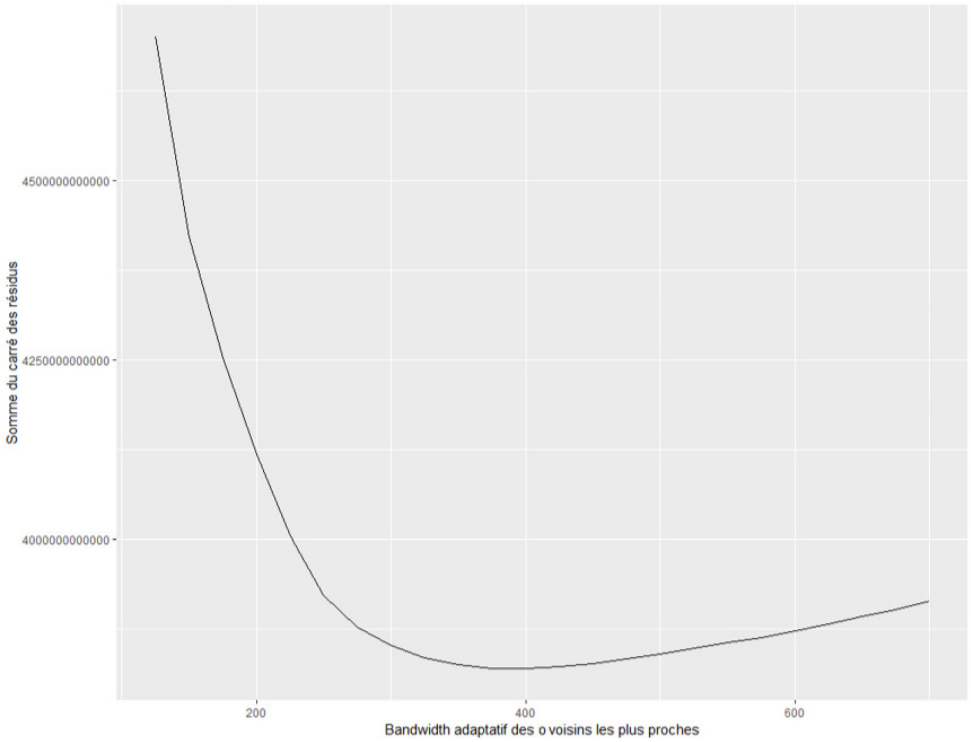
À cet effet, la figure 10 présente la distribution des valeurs de $o_{optimal}$ obtenues au sein des 100 échantillons d'apprentissage. Nous y constatons clairement que celles-ci varient, parfois substantiellement, d'un échantillon d'apprentissage à un autre.

Figure 10 - Distribution des valeurs de $o_{optimal}$ au sein des 100 échantillons d'apprentissage

$o_{optimal}$	Nombre d'échantillons d'apprentissage
350	2
375	52
400	28
425	14
450	2
475	2
Total	**100**

Par ailleurs, tel que nous l'avons évoqué aux sections 2.3 et 2.4, les indicateurs de performance des RGP sont également générés sur la base de l'ensemble des ventes, de manière à évaluer leur généralisation. À cette fin, nous calculons également $o_{optimal}$ sur la base de l'ensemble des 4 592 ventes. La figure 11 présente les résultats de la procédure CV en utilisant l'ensemble des ventes. Nous y constatons que $o_{optimal} = 400$ constitue la valeur qui minimise la somme du carré des résidus des RGP.

Figure 11 - Résultat de la procédure CV sur l'ensemble des ventes



2.5.4 MPH intégrant une segmentation *fuzzy* basée sur la proximité des propriétés (FUZZY)

Dans le cadre du présent mémoire, la segmentation *fuzzy* est réalisée par l'algorithme *fuzzy k-means*, intégrant exclusivement les variables numériques standardisées des coordonnées géographiques MTM. Les coefficients d'appartenance *fuzzy* sont calculés, avec le logiciel R, par la fonction *fcm* de la librairie *ppclust*.

Le nombre de segments optimal à inclure dans l'algorithme *fuzzy k-means* est déterminé, par essais et erreurs, sur la base de la technique par validation croisée décrite à la section 2.3, et comportant 10 itérations. Ce nombre réduit d'itérations est justifié par la lenteur de cet algorithme pour converger vers une solution optimale. L'objectif de l'utilisation de cette technique est de sélectionner le nombre de segments qui

optimise un indicateur de performance sélectionné par l'analyste. Dans le cadre de la présente étude, puisque nous sommes en contexte d'évaluation municipale, nous cherchons à minimiser l'écart type relatif à la médiane. Or, dans le cas présent, nous concluons que cet indicateur est minimisé lorsque le nombre de segments k = 80.

2.5.5 MPH intégrant la procédure itérative de segmentation aléatoire (PISA)

La procédure itérative de segmentation aléatoire, la PISA, constitue une toute nouvelle approche de segmentation territoriale, développée dans le cadre du présent mémoire. Celle-ci combine le cadre rigoureux des MPH classiques et une technique de classification non supervisée : la segmentation par nuées dynamiques. Il est intéressant de noter que le principal point faible réputé de cet algorithme, à l'effet que sa solution est tributaire des points de départ, devient son plus grand avantage dans le cadre de la PISA. En effet, c'est expressément ce défaut qui lui permet d'accomplir la tâche souhaitée et qui la distingue : celle de générer des segments (ou « *clusters* ») aléatoires, regroupant essentiellement des propriétés connexes ou voisines. À cet effet, les seules variables utilisées dans les nuées dynamiques sont les variables numériques standardisées des coordonnées géographiques MTM.

La PISA se fonde sur la prémisse que les forces spatiales propres à un emplacement donné sont si nombreuses et complexes, qu'il est pour le moins improbable qu'une seule segmentation *a priori* puisse capter les subtilités contenues dans les données de manière optimale. La procédure itérative, suggérée par la PISA, vient en quelque sorte répliquer le raisonnement d'un analyste qui souhaiterait obtenir, via un MPH, plusieurs estimations *a priori* des prix hédoniques des emplacements, sur la base de découpages distincts du territoire, générés aléatoirement, et regroupant essentiellement des propriétés connexes ou voisines.

Par ailleurs, nous avons évoqué précédemment que le nombre limité, voire l'absence d'observations, pour une unité de voisinage donnée, incite souvent l'analyste

à revoir sa stratégie de segmentation territoriale *a priori* : c'est-à-dire qu'il peut choisir d'extrapoler des conclusions provenant d'autres unités de voisinage, ou encore de modifier le niveau d'agrégation pour ces unités de voisinage spécifiques. Or, la PISA élimine cette problématique, puisque l'estimation de la valeur d'un emplacement donné ne repose pas exclusivement sur la présence de ventes au sein de son unité de voisinage, mais plutôt sur l'ensemble des ventes environnantes, qu'elles soient ou non localisées dans la même unité de voisinage.

Également, nous soulignons que les variables communément incluses dans les MPH, pour estimer les prix implicites des emplacements, sont généralement très sensibles à la multicolinéarité, à l'autocorrélation spatiale et à un biais d'omission (Dubin et Sung, 1990). Pour contourner ces problématiques, la PISA vise à substituer, dans le MPH final, les nombreuses variables usuelles, exogènes et endogènes, mesurant la valeur de l'emplacement, par un indice de prix hédonique de localisation (IPHL) qui intègre implicitement la valeur contributive de celles-ci. L'IPHL agit, dans le MPH, essentiellement à titre de variable proxy, un concept que nous avons abordé à la sous-section 1.2.1, dans l'exemple de l'école primaire prisée. Cet indice est calculé sur une base *a priori*, pour chaque propriété à évaluer, en fonction d'une procédure itérative permettant à un MPH d'estimer, non pas une seule, mais bien plusieurs régressions, intégrant chacune un découpage territorial généré aléatoirement et ayant pour seul critère la proximité des propriétés.

Sur la base des explications qui précèdent, pour ce MPH intégrant la PISA, il convient de modifier quelque peu l'équation 2 présentée à la sous-section 1.2.1, en la suivante :

$$Y = X\beta + E_{IPHL(v,\,z)}\,\Omega_{IPHL} + \mu \qquad \text{(équation 6)}$$

où β est un vecteur contenant les prix hédoniques de chacune des caractéristiques observables X, Ω_{IPHL} est le prix hédonique attribuable aux indices de prix hédonique

de localisation estimés pour chaque emplacement E localisé aux coordonnées géographiques MTM v et z, et μ constitue le terme d'erreur aléatoire identiquement et indépendamment distribué.

Les étapes qui suivent résument le fonctionnement de la PISA :

Étape 1 : Tirage aléatoire de k propriétés parmi l'inventaire, c'est-à-dire l'ensemble des n_I propriétés qui feront l'objet d'une estimation par le modèle;

Étape 2 : Création de k segments par nuées dynamiques sur les variables standardisées des coordonnées géographiques MTM, à partir des centres initiaux sélectionnés à l'étape 1, et sauvegarde du segment d'appartenance pour chaque propriété de l'inventaire;

Étape 3 : Estimation du MPH dans lequel on intègre k – 1 variables binaires représentant les k segments générés à l'étape 2;

Étape 4 : Pour i = 1, …, n_I, sauvegarde de b_{0i} et de b_{ji} avec $j \in [1, \dots, k-1]$ tel que la propriété i fait partie du segment j.

On exécute ces 4 étapes m fois [l = 1, … , m].

Étape 5 : À la suite des m itérations, pour i = 1, …, n_I, on calcule :

$$L_i = \sum_{l=1}^{m} \frac{(b_{0_{i_l}} + b_{j_{i_l}})}{m}$$

où $b_{0_{i_l}}$ est le coefficient de la constante pour la propriété i à l'itération l, et $b_{j_{i_l}}$ est le coefficient du segment j de l'itération l contenant la maison i.

Étape 6 : Finalement, on calcule, pour chacune des propriétés de l'inventaire, l'indice de prix hédonique de localisation, désigné $IPHL_i$:

$$IPHL_i = \frac{(L_i - L_{min})}{(L_{max} - L_{min})}$$

où $L_{min} = \min\{L_1, L_2 \dots L_{n_I}\}$ et $L_{max} = \max\{L_1, L_2 \dots L_{n_I}\}$.

Dans le MPH final, l'IPHL se substitue aux variables spatiales omises dans la spécification initiale du MPH.

Or, puisqu'elle repose sur les nuées dynamiques, une approche de segmentation non supervisée, la PISA requiert de déterminer initialement le nombre de segments k souhaité à chaque itération l, ainsi que le nombre d'itérations m. Nous posons donc que l'IPHL de la propriété i dépend des paramètres k et m de la PISA, c'est-à-dire que $IPHL_i$ dépend de PISA(k, m).

Or, une importante question se dresse : comment déterminer le nombre optimal de segments et d'itérations à inclure dans la PISA? Tout d'abord, il importe de noter que le nombre d'itérations joue surtout un rôle dans la réplicabilité de la solution, c'est-à-dire la capacité de celle-ci à générer des IPHL similaires si on l'exécute plusieurs fois sur les mêmes observations. Dans le cadre de cette étude, sur la base de plusieurs tests, nous avons remarqué qu'un nombre de 100 itérations procure une bonne réplicabilité, tout en étant relativement rapide à exécuter. Quant au nombre de segments, il joue un rôle important dans la précision des IPHL. Tout d'abord, nous posons :

$$t = \frac{\text{nombre de ventes dans l'échantillon d'apprentissage}}{\text{k}}$$

Afin de déterminer le nombre de segments optimal à retenir dans la PISA, nous préconisons, ici encore, une démarche par essais et erreurs, sur la base de la technique de validation croisée décrite à la section 2.3. Dans le présent mémoire, nous avons testé plusieurs valeurs de k, tel que t est compris entre 5 et 20. Puisque nous sommes en contexte d'évaluation municipale, nous cherchons à minimiser l'écart type relatif à la médiane des rapports estimation / prix réel. Sur cette base, un nombre k = 350 est identifié comme étant optimal.

Voilà qui conclut le second chapitre portant sur la méthodologie de la présente étude. Nous proposons maintenant de passer à la présentation et l'analyse des résultats.

CHAPITRE 3 : PRÉSENTATION ET ANALYSE DES RÉSULTATS

Maintenant que nous avons défini notre démarche méthodologique et abordé les spécificités des modèles, il est maintenant temps de confronter ceux-ci aux données. À cette fin, il est d'usage de débuter par une analyse descriptive des données.

3.1 Analyse descriptive des données

À Laval, nous dénombrons 84 116 propriétés correspondant aux critères précités à la section 2.2, et faisant partie du jeu de données de l'inventaire. Ce nombre équivaut à 93,64% de l'ensemble des 89 827 résidences d'un logement de type indivis contenues sur le territoire de Laval[7]. Le jeu de données des ventes compte, pour sa part, quelque 4 592 observations de propriétés transigées, représentant 5,46% de l'ensemble des propriétés du jeu de données de l'inventaire.

Dans un premier temps, il convient de remarquer, à la figure 12, que les ex-villes Îles-Laval et Laval-sur-le-Lac ne sont aucunement représentées dans les deux jeux de données. Cette constatation s'explique par le fait qu'aucune résidence unifamiliale au sein de ces deux ex-villes n'est desservie à la fois par les services d'égouts et d'aqueduc, alors que nous avons exclu ces caractéristiques de la présente étude, pour des raisons détaillées à la section 2.2.

[7] Source : ligne 304 du sommaire du rôle de la ville de Laval en date du 5 juillet 2018.

Figure 12 - Répartition des observations par ex-ville et par jeu de données

	INVENTAIRE		VENTES	
Ex-ville	Fréquence	Proportion	Fréquence	Proportion
Auteuil	6877	8.1756%	412	8.9721%
Chomedey	15202	18.0727%	651	14.1768%
Duvernay	8264	9.8245%	430	9.3641%
Fabreville	12511	14.8735%	748	16.2892%
Îles-Laval	0	0.0000%	0	0.0000%
Laval-des-Rapides	4526	5.3807%	209	4.5514%
Laval-Ouest	2998	3.5641%	141	3.0706%
Laval-sur-le-Lac	0	0.0000%	0	0.0000%
Pont-Viau	2052	2.4395%	107	2.3301%
Ste-Dorothée	8259	9.8186%	533	11.6071%
Ste-Rose	9708	11.5412%	692	15.0697%
St-François	4073	4.8421%	184	4.0070%
St-Vincent-de-Paul	2686	3.1932%	127	2.7657%
Vimont	6960	8.2743%	358	7.7962%
	84116	100.0000%	4592	100.0000%

Dans un second temps, à la figure 13, nous pouvons remarquer que les trois types de propriétés les plus représentés dans l'inventaire sont respectivement : les résidences de type plain-pied détaché (43,5078%), les résidences de type à étages entiers détaché (23,7077%) et les résidences de type à étages entiers jumelé (12,3116%). Nous constatons également que, pour le parc immobilier résidentiel unifamilial de Laval, le lien physique dominant est le type détaché (80,4461%), tandis que le genre dominant est le plain-pied (47,3323%).

Figure 13 - Répartition des observations par genre de propriété et lien physique et par jeu de données

	INVENTAIRE		VENTES	
Genre de propriété	Fréquence	Proportion	Fréquence	Proportion
De plain-pied	39814	47.3323%	1880	40.9408%
À niveaux décalés	8817	10.4820%	418	9.1028%
À étage mansardé	2891	3.4369%	130	2.8310%
À étages entiers	32594	38.7489%	2164	47.1254%
	84116	100.0001%	4592	100.0000%
Lien physique				
Détaché	67668	80.4461%	3485	75.8929%
Jumelé	14026	16.6746%	896	19.5122%
En rangée 1 côté	1016	1.2079%	86	1.8728%
En rangée 2 côtés	1406	1.6715%	125	2.7221%
	84116	100.0000%	4592	100.0000%
Genre et lien physique				
De plain-pied détaché	36597	43.5078%	1703	37.0862%
De plain-pied jumelé	3098	3.6830%	171	3.7239%
De plain-pied en rangée 1 côté	59	0.0701%	3	0.0653%
De plain-pied en rangée 2 côtés	60	0.0713%	3	0.0653%
À niveaux décalés détaché	8458	10.0552%	394	8.5801%
À niveaux décalés jumelé	355	0.4220%	23	0.5009%
À niveaux décalés en rangée 1 côté	1	0.0012%	1	0.0218%
À niveaux décalés en rangée 2 côtés	3	0.0036%	0	0.0000%
À étage mansardé détaché	2671	3.1754%	118	2.5697%
À étage mansardé jumelé	217	0.2580%	11	0.2396%
À étage mansardé en rangée 2 côtés	3	0.0036%	1	0.0218%
À étages entiers détaché	19942	23.7077%	1270	27.6568%
À étages entiers jumelé	10356	12.3116%	691	15.0479%
À étages entiers en rangée 1 côté	956	1.1365%	82	1.7857%
À étages entiers en rangée 2 côtés	1340	1.5930%	121	2.6350%
	84116	100.0000%	4592	100.0000%

Dans un troisième temps, nous présentons, à la figure 14, les statistiques descriptives de certaines variables numériques clés : la superficie habitable, la superficie du terrain, l'âge apparent de la maison (pour une définition, se référer à la figure 9) et le prix de vente. Les trois premières sont sélectionnées en raison de leur importance reconnue dans la formation des prix immobiliers, mais également parce qu'il s'agit de trois caractéristiques universelles pour les propriétés sous étude. Quant

au prix de vente, il s'agit d'une variable incontournable puisque cette dernière constitue la variable dépendante de notre étude.

À première vue, nous ne repérons aucune valeur impossible. Néanmoins, nous remarquons que, pour chacune des trois variables explicatives numériques présentées, c'est-à-dire la superficie habitable, la superficie du terrain et l'âge apparent de la maison, il existe des valeurs minimales et / ou maximales de l'inventaire se situant à l'extérieur de la plage constatée dans l'échantillon des ventes. Bien qu'il soit possible, et généralement peu risqué, d'extrapoler les résultats de l'analyse de régression à des valeurs hors de cette plage, la fiabilité des résultats s'en trouverait nécessairement quelque peu réduite. Nous choisissons donc, dans la présente étude, d'appliquer les résultats d'évaluation exclusivement aux propriétés dont les caractéristiques sont comprises dans la plage du jeu de données des ventes. Un filtre additionnel est donc appliqué sur les données de l'inventaire pour ne retenir que les propriétés dont la superficie habitable se situe entre 43,50 m^2 et 413,60 m^2, la superficie du terrain entre 115,60 m^2 et 1 590,50 m^2 et l'âge apparent entre 0 et 69 ans. À la suite de ce dernier filtre, le jeu de données de l'inventaire contient 83 872 observations qui feront l'objet d'une estimation. Au total, 244 observations ont été retranchées du jeu de données de l'inventaire par ce filtre additionnel, c'est-à-dire à peine 0,29% de celui-ci. Évidemment, aucune vente n'a été retranchée par ce procédé.

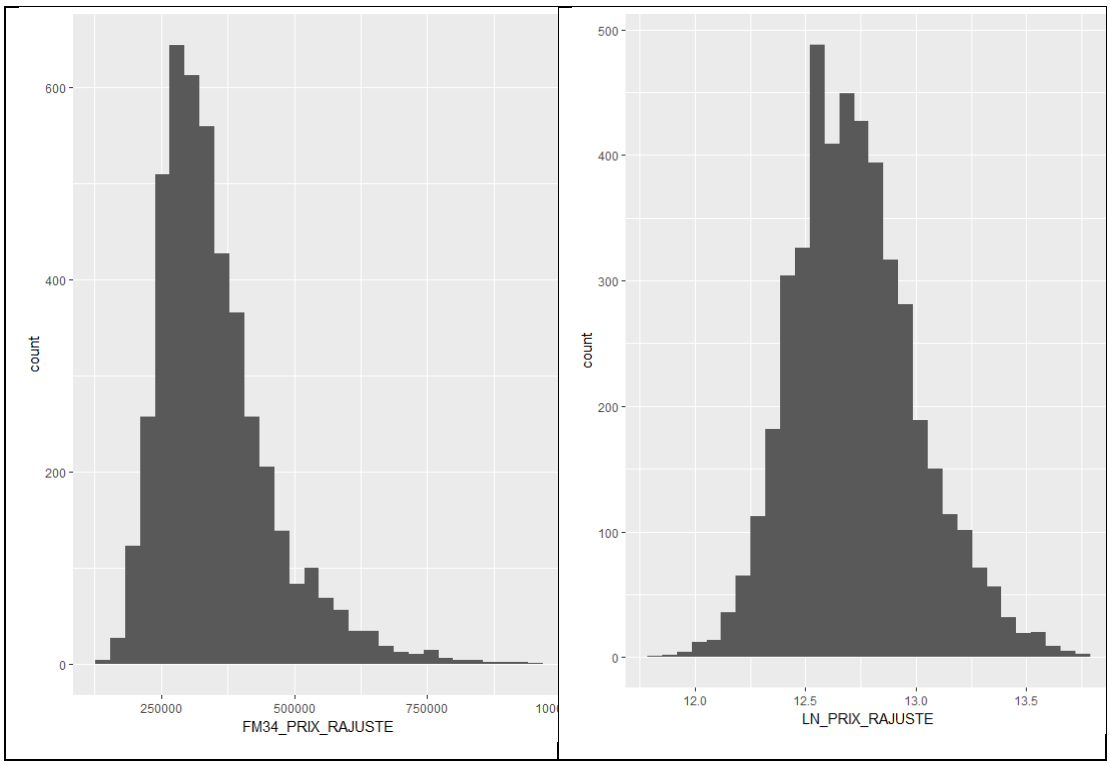
Figure 14 - Statistiques descriptives de certaines variables numériques clés

	INVENTAIRE				VENTES			
	Min	Max	Moy	Coef.var.	Min	Max	Moy	Coef.var.
Superficie habitable (m^2)	37,9	570,7	126,41	33,81%	43,5	413,6	128,19	33,16%
Superficie du terrain (m^2)	97,6	115 105,9	488,65	95,11%	115,6	1590,5	452,47	34,47%
Âge apparent en 2018	0	71	31,99	40,01%	0	69	28,03	48,37%
Prix de vente					137 000$	950 000$	351 027$	30,72%

Les figures 15 à 17 présentent les histogrammes, à la suite du retrait de ces 244 observations au sein de l'inventaire, des trois variables numériques continues précitées: c'est-à-dire le prix de vente, la superficie habitable et la superficie du terrain.

La figure 15 permet de constater, à gauche, la distribution de la variable dépendante non transformée, c'est-à-dire le prix de vente de la propriété en dollars. Nous détectons une asymétrie positive évidente, c'est-à-dire que la queue de la distribution s'étire vers la droite, témoignant de certains prix de vente extrêmes entraînant la moyenne à la hausse, par rapport aux autres indices de tendance centrale tels la médiane et le mode. Par ailleurs, du côté droit, nous visualisons clairement l'impact de la transformation logarithmique sur la distribution initiale des prix de vente : celle-ci est de toute évidence plus équilibrée et s'apparente davantage à une distribution normale. Les MCO étant sensibles aux valeurs extrêmes, à la fois du point de vue des coefficients que leur erreur type, la transformation logarithmique apparaît comme une option plus intéressante dans le cas présent.

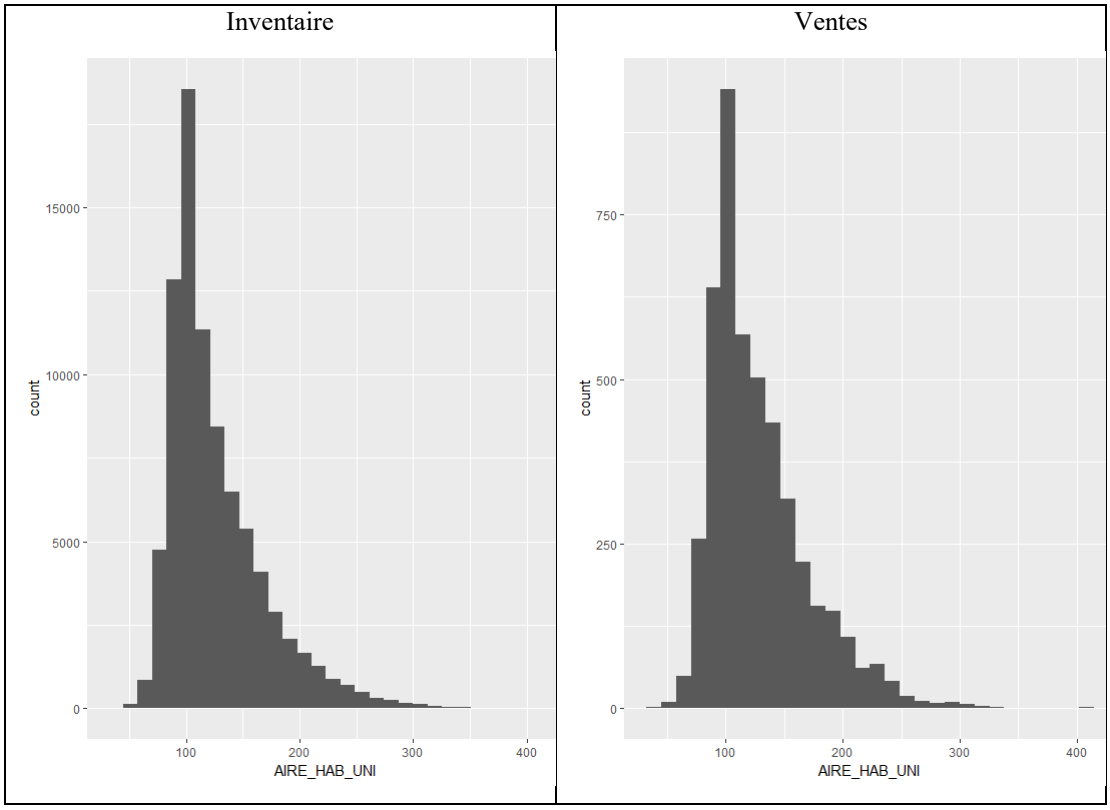
Figure 15 - Histogrammes de la variable non transformée du prix de vente (à gauche) et de la variable du prix de vente transformée par un logarithme népérien (à droite)



Nous rappelons ici que la forte étendue des prix de vente s'explique simplement par le fait qu'aucun filtre sur les variables explicatives identifiées précédemment ne permet de réduire celle-ci. À cet effet, nous suspectons l'influence d'une combinaison de variables explicatives incluses dans les modèles, dont l'influence de certains voisinages très prisés, dans la formation de ces prix de ventes extrêmes observés.

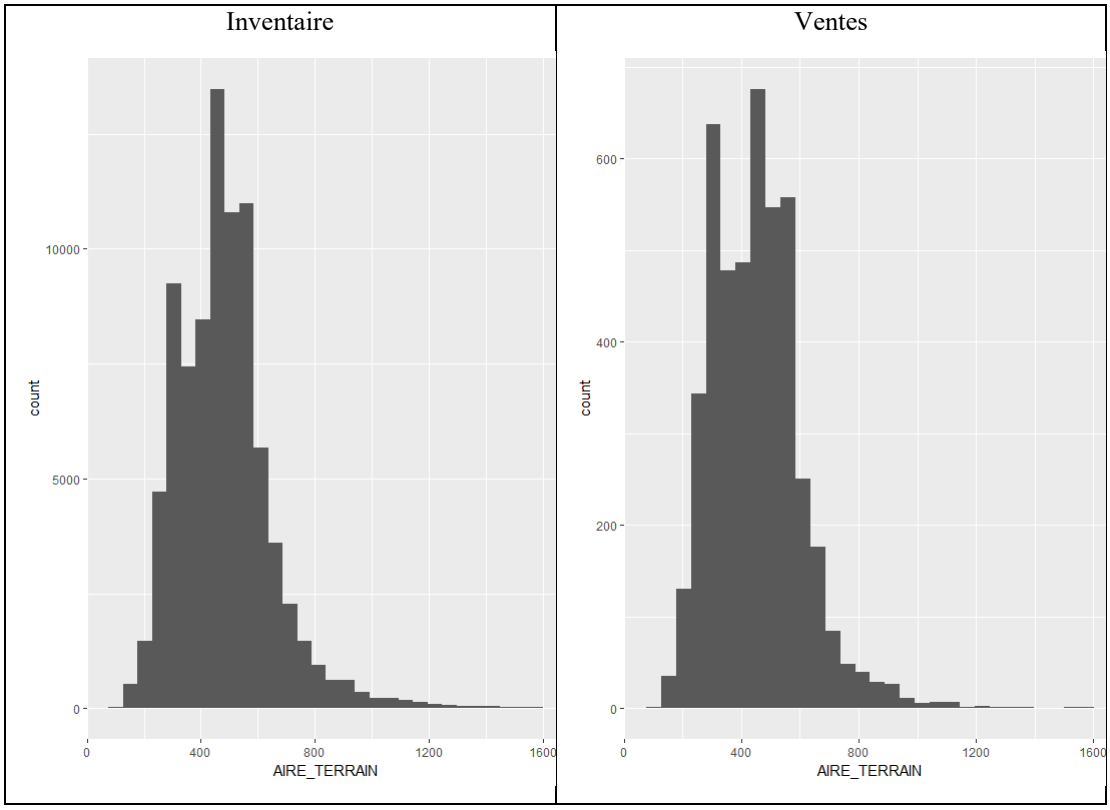
Quant à la figure 16, celle-ci présente les distributions relatives à l'aire habitable, à la fois pour l'inventaire et les ventes. Nous y détectons également une forte asymétrie positive. En évaluation immobilière, il convient de noter l'existence du principe de la contribution marginale décroissante, qui établit que la valeur contributive d'un attribut n'est pas nécessairement proportionnelle à sa quantité. Il en ressort qu'une transformation logarithmique de la variable de l'aire habitable peut s'avérer intéressante, ne serait-ce que pour tester la linéarité de la relation entre l'aire habitable d'une propriété et son prix de vente.

Figure 16 - Histogrammes de la variable de l'aire habitable pour l'inventaire et les ventes



À la figure 17, nous présentons les distributions relatives à la superficie du terrain, à la fois pour l'inventaire et les ventes. Nous y constatons également une forte asymétrie positive. Or, le principe de la contribution marginale décroissante nous incite donc, ici encore, à procéder à une transformation logarithmique pour cette variable explicative.

Figure 17 - Histogrammes de la variable de la superficie du terrain pour l'inventaire et les ventes



Pour conclure cette analyse descriptive des jeux de données, il est pertinent de présenter un portrait des découpages du territoire existant, élaborés par des évaluateurs professionnels, en conformité avec le règlement sur le rôle d'évaluation foncière. La figure 18 présente certaines statistiques les concernant. Nous y remarquons rapidement que les unités de voisinage sont le fruit d'un découpage très fin du territoire : nous comptons en moyenne à peine 100,21 résidences et 7,16 ventes par unité de voisinage. Par ailleurs, il importe de noter que nous dénombrons 196 unités de voisinage pour lesquelles nous ne répertorions aucune vente. Qui plus est, parmi les 641 unités de voisinage comportant au moins une vente, seulement 274 comprennent cinq ventes ou plus. C'est donc dire que seulement 32,74% des unités de voisinage présentes dans l'inventaire comportent au moins cinq ventes pour fins d'analyses. Pour leur part, les sous bassins d'analyses sont le fruit d'un découpage du territoire à plus haut niveau. Nous dénotons néanmoins que cinq sous bassins compris dans l'inventaire ne sont

aucunement représentés au sein des ventes. Une attention particulière sera portée à ces secteurs de l'inventaire, lors de l'étape de validation des résultats précédant le dépôt du rôle d'évaluation.

Figure 18 - Nombre et taille moyenne des unités de voisinage et sous bassins d'analyses

	INVENTAIRE	VENTES
Nombre d'unités de voisinage	837	641
Nombre moyen d'observations par unité de voisinage	100,21	7,16
Nombre de sous-bassins d'analyses	43	38
Nombre moyen d'observations par sous-bassin d'analyses	1 950,51	120,84

3.2 Résultats obtenus

La présente section aborde les résultats obtenus, pour chacun des modèles testés, selon la stratégie d'évaluation de la performance identifiée à la section 2.3.

3.2.1 Résultats obtenus par validation croisée

Nous avons fait valoir que les résultats obtenus, sur la base d'observations totalement indépendantes du processus de calibration des modèles, procure une meilleure évaluation de leur véritable performance sur des données qui n'ont pas encore été observées. Nous présentons donc ici les résultats permettant d'évaluer et comparer les modèles entre eux.

Indicateurs de performance : La figure 19 présente les résultats obtenus au niveau des indicateurs de performance sélectionnés, par la technique de validation croisée, c'est-à-dire sur des observations totalement indépendantes du processus de calibration des modèles.

Figure 19 - Indices de performance obtenus par la technique de validation croisée sur les indices de performance comportant 100 itérations

	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5
	NAIF	EXPERT	RGP	$PISA_{(350, 100)}$	$FUZZY_{(80)}$
I1. Proportion médiane des rapports estimation / prix réel	99,71%	99,99%	99,98%	99,90%	99,75%
I2. Écart type relatif à la médiane des rapports estimation / prix réel	10,55%	7,69%	8,10%	7,66%	7,96%
I3. Taux d'erreur moyen en valeur absolue (MAPE)	8,25%	5,99%	6,32%	5,98%	6,27%
I4. Erreur moyenne en dollars en valeur absolue (MAD)	28 334$	20 858$	21 957$	20 840$	22 127$
I5. % des taux d'erreur en valeur absolue inférieurs à 10%	67,76%	81,56%	79,89%	81,94%	79,57%
I6. Taux d'erreur moyen (MPE)	-0,58%	-0,35%	-0,43%	-0,29%	0,00%
I7. I de Moran	0.1323***	0,0070	0,0066	0,0072	0,0328**
*** Significatif au seuil de 0,1% ** Significatif au seuil de 1% * Significatif au seuil de 5%					

Performance prédictive des modèles : Tout d'abord, nous remarquons que les résultats avantagent le modèle PISA sur l'ensemble des indices mesurant la performance prédictive, c'est-à-dire les indices I2 à I5. Ce modèle fait une erreur moyenne de 5,98%, et de 20 840$, lorsqu'il estime les prix de vente. De plus, il estime 81,94% des prix avec un taux d'erreur en deçà de 10%. Finalement, il comporte un écart type relatif à la médiane des rapports estimation / prix de vente réel de 7,66%. En comparant ces indices de performance à ceux du modèle NAIF, c'est-à-dire celui n'intégrant aucunement la dimension spatiale, nous pouvons affirmer que la PISA contribue à diminuer le taux d'erreur moyen de l'ordre de 27,52%. Cette amélioration substantielle de la performance prédictive témoigne à la fois du rôle crucial de la localisation dans la formation des prix des résidences unifamiliales à Laval, mais également de l'efficacité de la PISA pour prendre en compte cette dimension importante. Une autre constatation intéressante se rapporte à la performance prédictive du modèle EXPERT, celui intégrant un découpage par des évaluateurs professionnels, qui s'est avérée la deuxième meilleure constatée dans la présente étude. À cet effet, le découpage par experts a engendré un taux d'erreur moyen de 5,99% et une erreur moyenne de 20 858$, soit respectivement à peine 0,01% et 18$ de plus que ceux générés par la PISA. Pour ainsi dire, nous pouvons qualifier les performances prédictives de ces deux modèles comme pratiquement équivalentes.

Biais des modèles : Par ailleurs, en ce qui a trait au biais des estimations, aucun modèle ne ressort comme un gagnant évident. Bien que l'indice I1 avantage le modèle EXPERT, son taux d'erreur moyen de -0,35% se classe au troisième rang, derrière les modèles FUZZY (0,00%) et PISA (-0,29%). La non convergence des indicateurs I1 et I6, pour les modèles 2 à 5, illustre l'importance de considérer plusieurs indicateurs pour mesurer un phénomène, à défaut de quoi nous aurions pu conclure à tort qu'un modèle est moins biaisé que les autres, et vice versa. La seule conclusion probante que nous pouvons tirer des indices I1 et I6, est que le modèle NAIF est celui pour lequel ces deux indices s'éloignent le plus de leur valeur optimale respective, c'est-à-dire un ratio médian des rapports estimation / prix de vente de 100%, et un taux d'erreur moyen de 0%. Cette constatation n'est sans doute pas étrangère à l'omission de la dimension de la localisation, que l'on sait être cruciale dans la formation des prix, et à l'autocorrélation spatiale des résidus qui en résulte.

Autocorrélation spatiale des résidus des modèles : En ce qui concerne l'AS des résidus, nous constatons que les modèles EXPERT, RGP et PISA s'avèrent tous les trois efficaces pour contrôler le phénomène. Leur statistique respective du I de Moran de 0,0070, 0,0066 et 0,0072 comporte chacune une p-value associée excédant 0,3822, faisant en sorte que nous ne rejetons pas, pour ces trois modèles, l'hypothèse nulle de l'indépendance spatiale à ce seuil. Le modèle NAIF ressort comme celui affichant la plus forte AS des résidus, avec un I de Moran de 0,1323 et une p-value associée de 0,000. Pour ce modèle, nous rejetons donc l'hypothèse de l'indépendance spatiale, au risque de se tromper une fois sur 1000. Également, nous concluons à une AS positive, en vertu de laquelle les résidus sont généralement entourés spatialement par des résidus de même signe. Ce résultat est totalement conforme aux attentes théoriques, puisque ce modèle omet de considérer la dimension spatiale, ayant pour effet d'engendrer de la dépendance spatiale. Quant au modèle FUZZY, nous concluons également au rejet de l'indépendance spatiale, mais au risque de se tromper une fois sur 100.

3.2.2 Résultats obtenus sur l'ensemble des ventes

Tel que mentionné précédemment, nous présentons également les indicateurs de performance obtenus sur la base de l'ensemble des ventes, de manière à évaluer la généralisation des modèles.

Indicateurs de performance : La figure 20 expose les résultats obtenus au niveau des indicateurs de performance, mais en utilisant l'ensemble des ventes pour calibrer les modèles.

Figure 20 - Indices de performance obtenus sur l'ensemble des ventes

	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5
	NAIF	EXPERT	$RGP_{(h=400)}$	$PISA_{(350, 100)}$	$FUZZY_{(80)}$
I1. Proportion médiane des rapports estimation / prix réel	99,73%	99,92%	100,04%	99,87%	99,91%
I2. Écart-type relatif à la médiane des rapports estimation / prix réel	10,43%	7,09%	6,62%	6,94%	7,74%
I3. Taux d'erreur moyen en valeur absolue (MAPE)	8,13%	5,53%	5,22%	5,44%	6,06%
I4. Erreur moyenne en dollars en valeur absolue (MAD)	27 992$	19 323$	18 088$	18 964$	21 203$
I5. % des taux d'erreur en valeur absolue inférieur à 10%	68,14%	84,43%	86,63%	85,46%	80,99%
I6. Taux d'erreur moyen (MPE)	-0,53%	-0,25%	-0,31%	-0,24%	-0,29%
I7. I de Moran	0.1371***	-0.0002	0.0048**	-0.0023	-0.0102***
I8. R^2	86,91%	93,79%	92,90%	93,99%	92,63%
I9. R^2 ajusté	86,82%	93,31%	92,37%	93,95%	92,45%
I10. Erreur type de la régression	0,1025	0,0730	0,0538	0,0690	0,0776
*** Significatif au seuil de 0,1% ** Significatif au seuil de 1% * Significatif au seuil de 5%					

Détérioration de la performance prédictive des modèles : Dans un premier temps, en comparant les figures 19 et 20, nous pouvons remarquer que tous les modèles intégrant une segmentation territoriale, lorsqu'ils sont calibrés sur l'ensemble des ventes, affichent une performance prédictive qui se veut surestimée, par rapport à celle obtenue sur des ventes indépendantes du processus de calibration. Ces résultats mettent en lumière l'importance de considérer la généralisation des modèles afin d'en évaluer de manière plus fiable la performance prédictive.

Or, nous constatons que cette baisse de la performance prédictive est d'autant plus évidente pour le modèle RGP, qui affiche la pire généralisation parmi les cinq modèles testés. À cet effet, son taux d'erreur moyen passe de 5,22% sur l'ensemble des ventes, à 6,32% par la technique de validation croisée, ce qui correspond à une détérioration du pouvoir prédictif de l'ordre de 21,07% selon cet indicateur. Quant à l'écart type relatif à la médiane, celui-ci passe de 6,62% à 8,10%, ce qui signale une détérioration de la performance prédictive de l'ordre de 22,36%. Dans notre étude, nous attribuons cette piètre généralisation des RGP à deux principaux facteurs. Le premier, et le plus important, relève du fait que les RGP utilisent et surpondèrent le sujet qui s'est transigé. Nous rappelons à cet effet que seulement 5,48% des propriétés de l'inventaire, à la suite du dernier filtre appliqué, ont fait l'objet d'une transaction, tandis que les RGP, si elles utilisent l'ensemble des ventes, bénéficient du sujet qui s'est transigé dans 100% des estimations. Or, la surpondération du sujet qui s'est transigé au détriment des autres observations confère aux RGP un avantage prédictif non négligeable, par rapport aux autres propriétés à évaluer qui ne disposent pas d'une telle observation privilégiée. À plus forte raison, nous constatons que cet avantage prédictif se trouve amplifié si le sujet comporte une ou plusieurs caractéristiques rares au sein de son rayon, c'est-à-dire parmi les observations retenues pour estimer sa régression locale. En pareil cas, cette dernière attribue une partie non négligeable, voire la totalité, des fluctuations dans les prix à cette ou ces caractéristiques rares, ayant pour effet direct de surestimer la performance prédictive. La figure 21 présente un exemple d'une telle situation. Celui-ci se rapporte à la seule propriété localisée en zone inondable 0-20 ans, au sein de sa régression locale. Il n'est donc pas surprenant que les RGP affichent une erreur nulle pour cette observation spécifique. Or, nous constatons que l'estimateur local relatif à la variable indicatrice de la localisation en zone inondable 0-20 ans est de 0,029, ce qui est contraire à la logique. À plus forte raison, cette erreur nulle des RGP détourne notre attention sur une potentielle variable omise : à savoir la proximité de la rivière des Prairies, au fond de la rue.

Figure 21 - L'utilisation de l'observation du sujet, comportant une ou plusieurs caractéristiques rares, au sein de sa régression locale, a pour effet de surestimer la performance prédictive des RGP

Source : Google StreetView

Par ailleurs, un second facteur identifié, moins important, expliquant la détérioration de la performance prédictive des RGP concerne le processus de détermination du rayon optimal, lorsque celui-ci inclut les propriétés à évaluer. À cet effet, nous désirons rappeler un résultat obtenu dans la présente étude (figure 11), qui indique que divers sous-échantillons d'apprentissage pointent vers des rayons optimaux distincts. En pratique, l'analyste ne connaît pas le rayon optimal des propriétés non transigées qu'il souhaite évaluer. Il peut seulement déduire celui-ci à partir de transactions avérées. Ce processus a été simulé dans notre étude par la technique de validation croisée.

Coefficients des modèles : Dans un autre ordre d'idée, la figure 22 présente les coefficients générés pour chaque modèle sur la base de régressions utilisant l'ensemble des ventes. La présentation des coefficients des RGP diffère de celle des quatre autres modèles, étant donné leur caractère local et du fait, qu'en théorie, il peut y avoir autant de coefficients différents pour une variable donnée qu'il y a de points de régression.

Figure 22 - Coefficients et degrés de signification générés en utilisant l'ensemble des 4 592 ventes

	MODÈLE 1		MODÈLE 2		MODÈLE 3			MODÈLE 4		MODÈLE 5	
	NAIF		EXPERT		RGP			PISA		FUZZY	
	Coefficient	Sig	Coefficient	Sig	Min	Max	Moy	Coefficient	Sig	Coefficient	Sig
CONSTANTE	10,14577	***	10,73241	***	9,17461	12,50348	10,79200	10,48995	***	10,61039	***
AGE_APP_2018	-0,00635	***	-0,00912	***	-0,01241	-0,00538	-0,00916	-0,00916	***	-0,00906	***
AIRE_GAR_ATT_INT	0,00247	***	0,00187	***	0,00048	0,00402	0,00199	0,00175	***	0,00200	***
AIRE_GAR_SSOL	0,00184	***	0,00099	***	-0,00093	0,00282	0,00110	0,00090	***	0,00107	***
AIRE_HAB_UNI	0,00013		0,00077	***	-0,00179	0,00415	0,00088	0,00067	***	0,00080	***
AIRE_HAB_UNI_LN	0,40309	***	0,23094	***	-0,17518	0,60128	0,22313	0,24651	***	0,23973	***
AIRE_SSOL_FINIE	0,00053	***	0,00055	***	-0,00018	0,00151	0,00055	0,00053	***	0,00053	***
IF_1_SDB	-0,02792	***	-0,02003	***	-0,05323	0,01196	-0,02302	-0,02054	***	-0,02209	***
IF_ASP_CENTRAL	0,00518		0,00941	***	-0,02305	0,03670	0,00988	0,00954	***	0,01105	***
IF_BUNGALOW	-0,00480		0,00409		-0,09209	0,09118	-0,00500	0,00628		0,00246	
IF_CLIM_CENTRAL	0,05435	***	0,03001	***	0,00366	0,08266	0,03405	0,02985	***	0,03336	***
IF_CLIM_MURAL	0,01741	***	0,01361	***	-0,01276	0,04332	0,01401	0,01239	***	0,01259	***
IF_EN_RANGEE_1	-0,08271	***	-0,09878	***	-0,20464	0,06402	-0,08043	-0,09513	***	-0,09237	***
IF_EN_RANGEE_2	-0,13462	***	-0,11779	***	-0,31500	0,02928	-0,10625	-0,11184	***	-0,11320	***
IF_FOND_PAS_SSOL	-0,07076	***	-0,03770	***	-0,18014	0,07008	-0.04673	-0,04330	***	-0,05653	***
IF_FOYER	0,01664	***	0,01113	***	-0,01469	0,05367	0,01308	0,01126	***	0,01285	***
IF_INON_0_20	-0,13008	***	-0,04123	**	-0,18111	0,22253	-0,01112	-0,02030	*	-0,07454	***
IF_INON_20_100	-0,07938	***	-0,03073	**	-0,20987	0,11715	-0,02210	-0,01215		-0,04192	***
IF_JUMELE	-0,07564	***	-0,07220	***	-0,16705	-0,01537	-0,08282	-0,08391	***	-0,08300	***
IF_MANSARDE	-0,02871	**	-0,00061		-0,21657	0,11547	-0,01159	-0,00221		-0,00782	
IF_NIV_DECALES	0,00007		-0,00585		-0,08118	0,08069	-0,00462	0,00033		-0,00131	
IF_PISCINE_EXC	0,04825	***	0,04177	***	-0,00400	0,09864	0,04786	0,04225	***	0,04802	***
IF_PISCINE_HT	-0,00139		0,00867	**	-0,03993	0,05608	0,00666	0,00949	***	0,01067	***
IF_VENTE_2016	-0,02996	***	-0,03087	***	-0,06195	-0.00350	-0,03090	-0,03103	***	-0,03157	***
IF_VENTE_2018	0,04360	***	0,03960	***	-0.09528	0,16690	0,03858	0,03808	***	0,03571	***
POINTAGE_CLASSE	0,00225	***	0,00170	***	-0,00035	0,00503	0,00188	0,00173	***	0,00187	***
POURC_BRIQUE_PIERRE	0,00048	***	0,00025	***	-0,00091	0,00121	0,00027	0,00026	***	0,00026	***
POURC_PLANCHER_SUP	0,00042	***	0,00020	***	-0,00064	0,00079	0,00019	0,00022	***	0,00021	***
POURC_VINYLE	-0,00042	***	-0,00018	***	-0,00131	0,00032	-0,00031	-0,00023	***	-0,00028	***
SUP_TERRAIN_LN	0,08108	***	0,12100	***	0,03466	0,23490	0,12580	0,12050	***	0,11788	***
SB_ANALYSES (31 variables)				***							
UV_ANALYSES (274 variables)				***							
IPHL (1 variable)								0,00051	***		
FUZZY (79 variables)											***

*** Significatif au seuil de 0,1%
** Significatif au seuil de 1%
* Significatif au seuil de 5%

Biais induit par l'omission de variables de localisation importantes : À titre de première constatation au niveau de l'analyse des coefficients, il importe de souligner l'influence des variables de localisation omises sur plusieurs estimateurs du modèle NAIF. Cette omission se manifeste par des coefficients biaisés, qui divergent de manière significative par rapport à ceux des quatre modèles intégrant la dimension spatiale.

En guise d'exemple, la figure 23 présente les valeurs contributives, fournies par chaque modèle, pour un garage attaché et un garage au sous-sol typique de 22,3 mètres carrés, et pour une maison hypothétique valant 350 000$. Nous remarquons rapidement

que les valeurs contributives fournies par le modèle NAIF, à savoir 19 800$ pour un garage attaché de 22,3 mètres carrés et 14 700$ pour un garage au sous-sol de même superficie, excèdent largement celles des quatre autres modèles, qui en moyenne affichent des valeurs respectives de 15 200$ et 8 000$. Qui plus est, selon le modèle NAIF, un garage attaché vaut environ seulement 35% de plus qu'un garage de même superficie au sous-sol, tandis que cette plus-value grimpe à 90% si l'on se fie aux quatre modèles intégrant une segmentation territoriale. Pour expliquer ce phénomène, il convient de noter que les deux variables explicatives relatives aux garages attachés ou au sous-sol, se veulent positivement et significativement corrélées avec la variable IPHL, mesurant le prix hédonique de l'emplacement. C'est donc dire que les secteurs les plus en demande, à Laval, comportent davantage de maisons avec garages. L'omission de considérer la dimension de la localisation, dans le modèle NAIF, attribue donc, de manière indue, la plus-value attribuable à la désirabilité du secteur, à certaines variables explicatives autocorrélées spatialement, dont les deux variables précitées. Cet exemple sert à illustrer le biais induit dans les estimateurs en omettant de considérer adéquatement la dimension spatiale dans les MPH.

Figure 23 - Valeurs contributives de garages, attaché et au sous-sol, de 22,3 mètres carrés en fonction d'une maison hypothétique valant 350 000$

	MODÈLE 1	**MODÈLE 2**	**MODÈLE 3**			**MODÈLE 4**	**MODÈLE 5**
	NAIF	**EXPERT**	**RGP**			**PISA**	**FUZZY**
	Moy.	Moy.	Min.	Max.	Moy.	Moy.	Moy.
Garage attaché	$19,800	$14,900	$3,800	$32,800	$15,900	$13,900	$16,000
Garage au sous-sol	$14,700	$7,800	-$7,200	$22,700	$8,700	$7,100	$8,500

Comparaison des coefficients des modèles : Dans un autre ordre d'idée, il est loisible de comparer les coefficients générés par les deux modèles les plus performants au niveau de la performance prédictive : les modèles PISA et EXPERT. Une première constatation est à l'effet que les coefficients générés par ceux-ci se ressemblent de manière générale, sans toutefois être équivalents. Par ailleurs, nous remarquons qu'au sein de ces deux modèles, la quasi-totalité des estimateurs sont significatifs au seuil de 0,10%, exceptées les variables reliées au genre de propriété (bungalow, à étages entiers,

à étages mansardés et à niveaux décalés). Nous pouvons donc conclure, pour les propriétés sous étude et en fonction des spécifications respectives retenues par ces modèles, que le genre de propriété n'exerce aucun effet, *ceteris paribus*, dans la formation des prix. Finalement, il importe de noter que tous les coefficients, pour ces deux modèles, affichent des résultats conformes aux attentes, en termes de signes et de valeurs. En guise d'exemple, d'un point de vue à la fois théorique et pratique, le type de lien physique d'une propriété influence sa désirabilité et sa rareté, donc sa valeur. Or, nous percevons clairement, dans les modèles EXPERT et PISA, que l'ordonnancement des valeurs contributives selon le type de lien physique respecte cette logique : les résidences détachées comportent une valeur qui, *ceteris paribus*, excède celle des jumelés de l'ordre de 6,97% (EXPERT) et 8,05% (PISA), ainsi que celle des maisons de ville à un seul voisin mitoyen à raison de 9,41% (EXPERT) et 9,07% (PISA), et finalement celles des maisons de ville comportant deux voisins mitoyens à raison de 11,11% (EXPERT) et 10,58% (PISA). Les coefficients obtenus sont donc totalement conformes aux attentes puisque, *ceteris paribus*, une résidence détachée est préférée à une résidence jumelée, qui se veut elle-même plus appréciée qu'une maison de ville comportant un seul voisin mitoyen, qui a son tour est préférée à une maison de ville comportant deux voisins mitoyens.

Pour conclure cette analyse des coefficients, il convient de commenter les résultats obtenus par les RGP. Tout d'abord, nous remarquons que les coefficients moyens générés par ce modèle sont, pour la plupart, semblables à ceux des modèles EXPERT et PISA, mais exhibent certains résultats moins convaincants au niveau de l'interprétation. Pour reprendre l'exemple du type de lien physique, et contrairement aux deux autres modèles précités, les estimateurs des RGP favorisent légèrement, *ceteris paribus*, une maison de ville comportant un seul voisin mitoyen à une maison de type jumelé. Bien que l'écart entre les deux est peu prononcé, il n'en demeure pas moins que ce résultat, d'un point de vue théorique et pratique, est pour le moins contestable. À cet effet, en analysant les coefficients minimums et maximums, on constate au sein des RGP, plusieurs coefficients locaux de signes opposés aux attentes

théoriques. Par exemple, il est pour le moins surprenant que la présence d'une seule salle de bains dans une résidence ajoute, *ceteris paribus*, $e^{0,01196} - 1 = 1,20\%$ à la valeur d'une propriété, peu importe le secteur. Dans le même ordre d'idée, il est inconcevable que des maisons de ville à un voisin (0,06402) et à deux voisins (0,02928) comportent une valeur plus élevée, *ceteris paribus*, qu'une résidence détachée. Un dernier exemple concerne les variables indicatrices de zone inondable de 0-20 ans et de 20-100 ans. À cet effet, les coefficients maximums positifs respectifs de 0,22253 et 0,11715 ne font aucun sens d'un point de vue de l'évaluation immobilière, puisque la localisation en zone inondable ne peut être positive, *ceteris paribus*, dans la formation des prix immobiliers. Cette constatation n'est pas sans rappeler la problématique évoquée à la figure 21.

Quant au modèle FUZZY, celui-ci présente des estimateurs totalement conformes aux attentes, en termes de signe et de valeur.

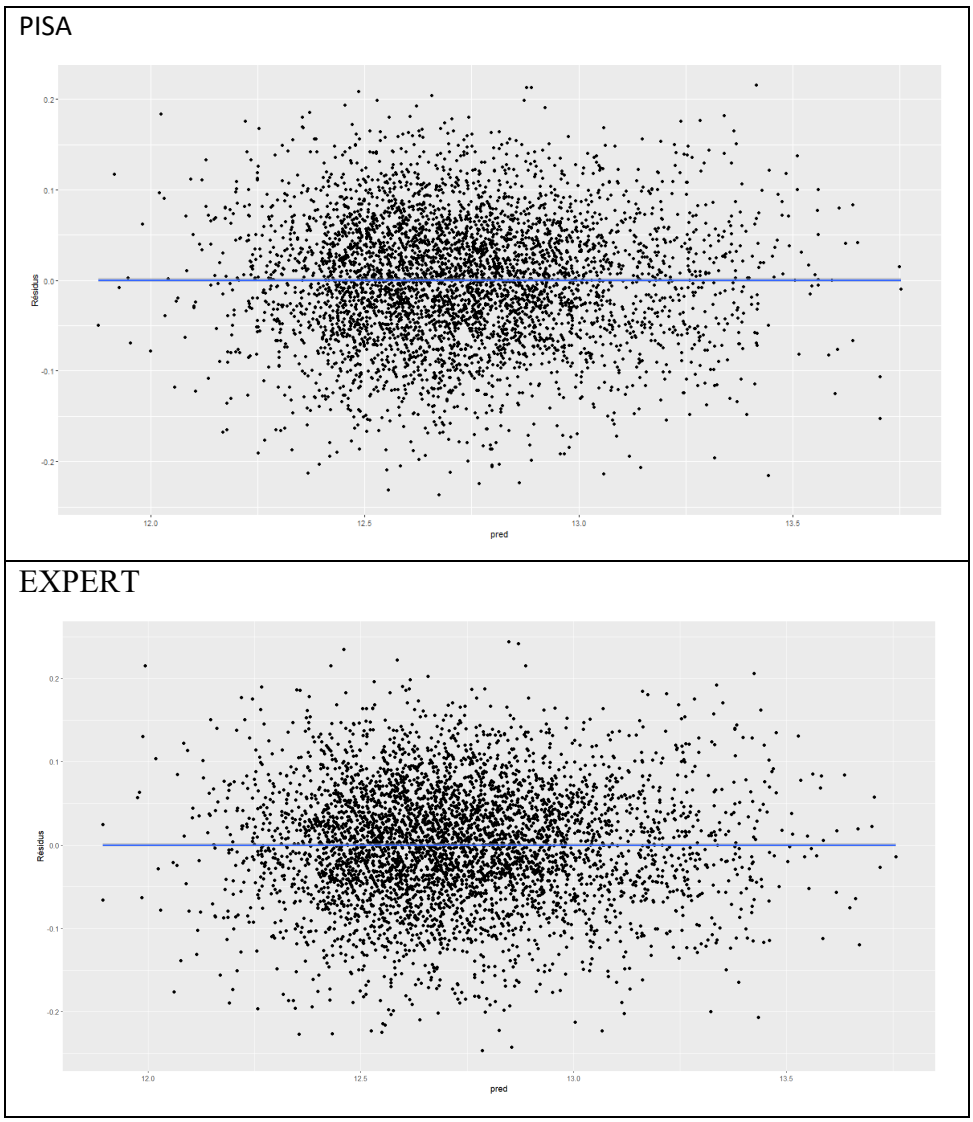
Distribution des résidus des modèles PISA et EXPERT : Bien que l'objectif premier de ce mémoire soit d'optimiser la performance prédictive, nous devons néanmoins nous assurer que les modèles développés comportent une bonne validité. À cet effet, la distribution des résidus donne beaucoup d'informations sur celle-ci. Nous évaluons donc la distribution des résidus des deux modèles s'étant démarqués au niveau de la performance prédictive, c'est-à-dire les modèles PISA et EXPERT.

Il existe de nombreux tests formels pour évaluer la spécification d'un modèle (test RESET de Ramsey), l'hétéroscédasticité (test de White) et la normalité des résidus (test de Jarque Bera). Toutefois, dans le cadre de la présente étude, nous préconisons des outils graphiques, puisque ces derniers permettent non seulement de détecter ces problèmes, mais également d'en visualiser la provenance.

La figure 24 présente les nuages de points des résidus en fonction des estimations (abscisses) des modèles PISA (en haut) et EXPERT (en bas). Nous

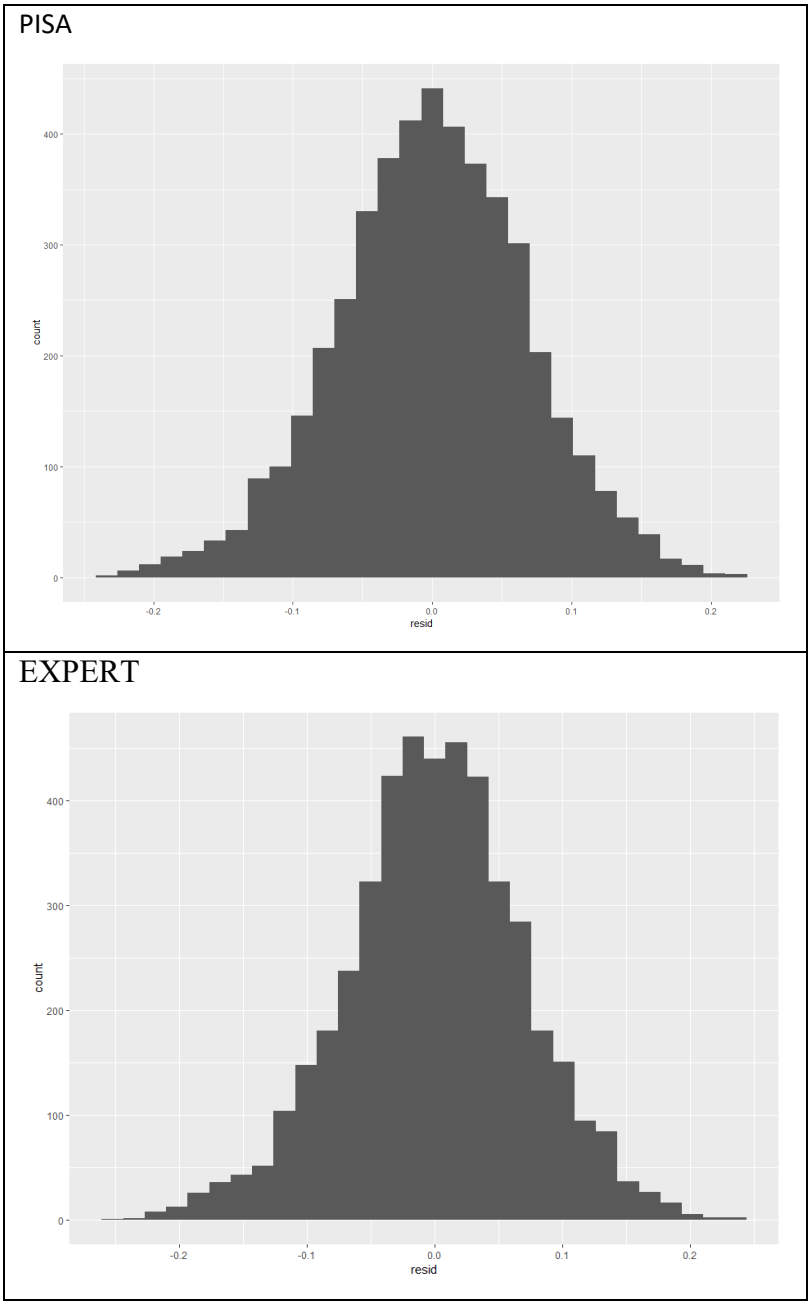
pouvons remarquer, pour les deux modèles, que les résidus semblent distribués aléatoirement autour de la droite horizontale représentant l'erreur nulle. Nous ne détectons aucune forme particulière dans ceux-ci qui incitent à suspecter la présence de liens non linéaires non modélisés. Également, les variances des résidus semblent relativement constantes selon les différentes valeurs des estimations. Les résidus aberrants sont absents, en ce sens que tous comportent une valeur inférieure à 0,25 en forme semi-log, ce qui est loin d'être problématique, à tout le moins pour le créneau sous étude.

Figure 24 - Nuages de points des résidus en fonction des estimations des modèles PISA et EXPERT



Quant à la figure 25, celle-ci affiche les histogrammes des résidus pour les modèles PISA et EXPERT. Nous constatons rapidement que ceux-ci présentent des distributions s'apparentant à une loi normale.

Figure 25 - Histogrammes des résidus des modèles PISA et EXPERT



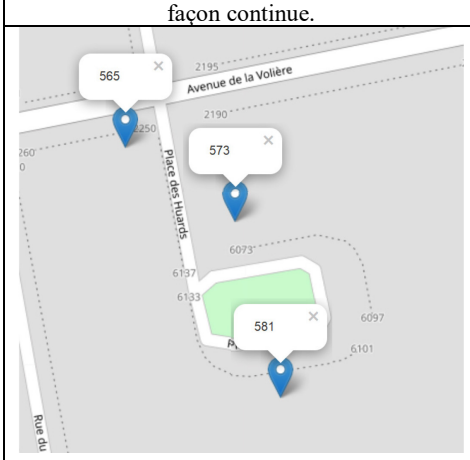
Sur la base de cette analyse sommaire des résidus, il appert que les modèles PISA et EXPERT comportent une bonne validité.

CHAPITRE 4 : DISCUSSIONS

Dans un premier temps, nous revenons sur les principaux objectifs de la présente étude. Le premier était de comparer les performances prédictives de diverses segmentations territoriales dans les MPH, en vue d'une utilisation pratique pour le rôle d'évaluation triennal de 2019-2020-2021 à la ville de Laval. À cet effet, les résultats pointent vers deux grands gagnants : à savoir le modèle PISA qui remporte la palme, talonné de très près par le modèle EXPERT. À l'ombre des résultats obtenus, nous concluons que la segmentation par unités de voisinage, telle que prescrite par le Manuel d'évaluation foncière, se veut une approche performante, à la fois en en termes de performance prédictive et de contrôle de l'autocorrélation spatiale des résidus. Qui plus est, le concept de voisinage se veut intuitif et généralement bien compris et accepté par les contribuables.

Néanmoins, sur la base des résultats obtenus, le paradigme à l'effet que l'influence de la localisation dans la formation des prix se manifeste strictement de manière discrète, c'est-à-dire au sein de voisinages observables et délimités, peut être contesté. À cet effet, la meilleure performance obtenue par la PISA, qui préconise une mesure continue de la valeur de l'emplacement, suggère que cette dernière obéit à des forces qui se veulent au moins aussi continues que discrètes. La figure 26 illustre cette divergence de paradigme. Nous y constatons, à gauche, que la PISA génère des IPHL plus élevés, sur la rue Place des Huards, à mesure que l'on s'éloigne de l'Avenue de la Volière, qui constitue une voie relativement achalandée. À droite, nous remarquons que l'entièreté de la rue Place des Huards est incluse dans une seule et même unité de voisinage, et que celle-ci englobe un tronçon de l'avenue de la Volière.

Figure 26 - Divergence de paradigme : L'influence de la localisation se manifeste-t-elle de manière continue ou discrète dans la formation des prix?

PISA	Unités de voisinage
L'influence de la localisation se manifeste de façon continue.	L'influence de la localisation se manifeste de façon discrète.
	

Or, les performances prédictives quasi-équivalentes obtenues par ces deux segmentations, pourtant totalement différentes, amènent une intéressante question : se pourrait-il que l'influence de la localisation se manifeste à la fois de façon continue et discrète dans la formation des prix? La PISA, du fait qu'elle procure une mesure continue de la valeur des emplacements, apparaît comme une meilleure option dans les secteurs caractérisés surtout par des forces spatiales continues. Pour sa part, le découpage *a priori* apparaît comme tout indiqué dans les secteurs où interviennent surtout des forces discrètes. La figure 27 présente un exemple appuyant cette hypothèse. Ce graphique a été généré en calculant préalablement les taux d'erreur moyens des modèles PISA et EXPERT pour chacune des unités de voisinage, sur la base de l'ensemble des 4 592 ventes. Les deux couleurs identifient le modèle qui se veut en moyenne le plus précis, pour une unité de voisinage donnée. Ce graphique parle de lui-même. D'une part, il permet de valider la présence de segments spatiaux évidents, indiquant que la performance des deux modèles précités se veut autocorrélée spatialement. D'autre part, ce graphique laisse entrevoir que la PISA est plus précise sur la majeure partie du territoire. À cet effet, nous notons que le modèle PISA surpasse

le modèle EXPERT, en termes de taux d'erreur moyen, dans 52,11% des unités de voisinage, lesquelles regroupent 58,97% des propriétés.

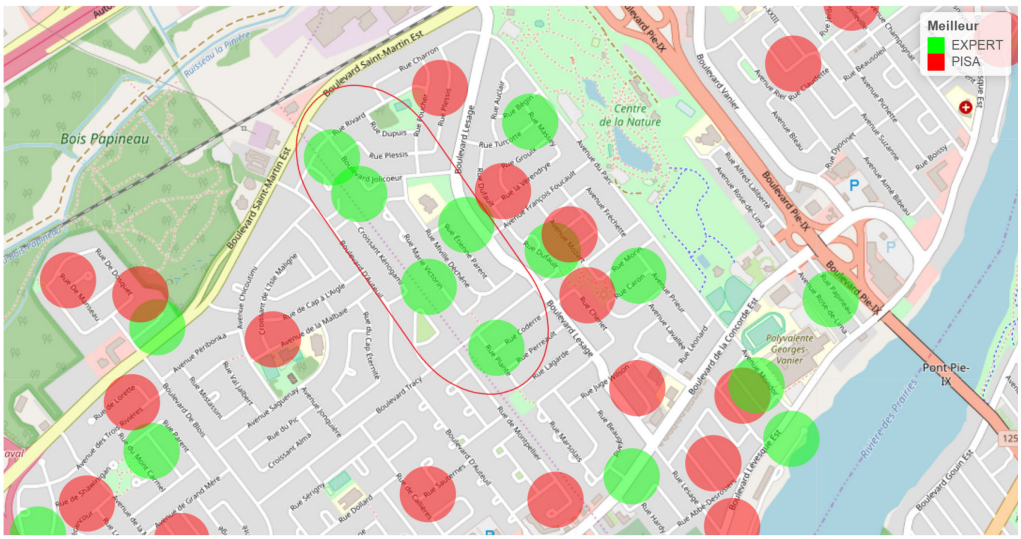
Figure 27 – Meilleurs taux d'erreur moyens générés par les modèles PISA et EXPERT par unité de voisinage : la performance des modèles est autocorrélée spatialement

Or, si la segmentation par unités de voisinage omet potentiellement de considérer certaines forces continues pertinentes, la PISA comporte aussi certaines limites. Une de celles-ci est représentée à la figure 28 et se rapporte à la nécessité de bien identifier les forces de localisation contribuant à la valeur de manière discrète sur une base *a priori*. Nous remarquons, dans l'image de gauche, que les IPHL générés sur le boulevard Jolicoeur (moyennes de 748,5 et 816,83), une voie séparant deux secteurs commandant des prix d'emplacement drastiquement différents, viennent capter l'influence du secteur adjacent très prisé localisé au sud-ouest (IPHL moyen de 843,2). L'omission de considérer cette force discrète a pour effet de biaiser localement l'IPHL, et de surévaluer du même coup toutes les propriétés sur ce boulevard. Dans l'image de droite, on constate que les unités de voisinage démontrent un découpage fin et particulier captant ces forces.

Figure 28 - Une limite de la PISA : l'omission de considérer des variables discrètes importantes biaise l'IPHL sur une base locale

La figure 29 permet justement de constater que le modèle EXPERT performe mieux que la PISA en présence de forces discontinues significatives. Nous y constatons, dans le même secteur que la figure 28, que les cinq unités de voisinage se trouvant à la jonction de cette coupure entre ces deux secteurs très différents, favorisent toutes le modèle EXPERT.

Figure 29 - Le modèle EXPERT performe mieux en présence de forces discrètes



Dans un autre ordre d'idée, le second objectif de ce mémoire était de valider si le découpage *a priori* du territoire sur la base de connaissances préalables quant aux forces et facteurs du marché s'y appliquant, se veut indispensable pour optimiser la performance prédictive des MPH. Les résultats obtenus dans la présente étude indiquent que la PISA, une approche de segmentation territoriale centrée sur les données, s'avère la technique individuelle la plus performante pour optimiser la performance prédictive. Puisqu'elle surpasse le découpage par experts, nous pouvons conclure qu'un découpage *a priori* sur la base des connaissances du territoire n'est pas formellement requis pour générer d'excellentes performances prédictives dans les MPH. Toutefois, les résultats obtenus font ressortir que les forces spatiales se manifestent à la fois de manière continue et discrète. Nous concluons donc qu'une segmentation optimale, sur le territoire de Laval, implique à la fois un découpage *a priori*, délimitant les phénomènes discrets, et une mesure continue de la valeur de l'emplacement telle que celle fournie par la PISA. Si nous supposons que le découpage *a priori* est optimisé par l'intervention d'un expert, surtout dans les secteurs où l'on retrouve des forces significatives et peu d'observations, rien n'indique toutefois qu'un tel découpage ne pourrait être généré strictement sur la base des données. Cette piste demeure à explorer.

Par ailleurs, il a été évoqué à la sous-section 1.1.1.3, qu'en raison de la problématique de l'aire spatiale modifiable, l'effet *ceteris paribus* des estimateurs doit être interprété par rapport à un modèle donné, c'est-à-dire en compagnie des autres estimateurs, et non de manière isolée ou absolue. À cet effet, les résultats présentés aux figures 22 et 23 illustrent, de manière non équivoque, que les coefficients des 29 variables explicatives communes aux cinq modèles varient, parfois substantiellement, selon l'approche de segmentation du territoire retenue. Cette considération est particulièrement importante lorsqu'il s'agit d'utiliser les estimateurs générés par les MPH pour ajuster les prix de vente des comparables dans l'application de la technique des prix de vente ajustés. Les estimateurs d'un MPH se doivent donc d'être utilisés

conjointement, avec l'ensemble des autres estimateurs de ce même modèle, sans quoi ils peuvent compromettre les résultats et induire en erreur.

L'originalité de la présente étude se décline sous plusieurs angles. D'une part, elle se démarque par son application pratique dans le domaine de l'évaluation municipale : les résultats obtenus guideront la conception du rôle d'évaluation triennal de 2019-2020-2021 à la ville de Laval. D'autre part, le présent mémoire inaugure une toute nouvelle approche pour intégrer la dimension de la localisation dans les MPH, la PISA, qui a généré la meilleure performance prédictive parmi les cinq modèles testés. Cette nouvelle technique de segmentation est la seule préconisant un découpage aléatoire du territoire, sur la stricte base de la proximité des propriétés, s'appuyant sur la prémisse qu'il existe non pas un seul, mais bien plusieurs découpages *a priori* admissibles. La procédure itérative, suggérée par la PISA, vient en quelque sorte répliquer le raisonnement d'un analyste qui souhaiterait obtenir plusieurs estimations de la valeur d'un emplacement, sur la base de découpages distincts du territoire, générés aléatoirement et regroupant essentiellement des propriétés connexes ou voisines. Aussi, nous n'avons répertorié aucune autre étude en évaluation immobilière distinguant les propriétés à évaluer, c'est-à-dire l'inventaire, de celles ayant été transigées, c'est-à-dire les ventes. Une telle manière de procéder assure l'applicabilité des résultats à la population visée.

Une force de ce mémoire est d'évaluer la performance des modèles par une technique de validation croisée, qui tient à la fois compte de la propension des modèles au surapprentissage, et qui assure que la performance mesurée n'est pas uniquement le fruit de la composition d'un échantillon donné. Une telle approche a notamment permis de faire ressortir une piètre généralisation de la part des RGP. Une autre force se rapporte au grand nombre d'observations (n_v = 4 592), et à la qualité des données qui s'est avérée excellente.

Quant aux faiblesses, on peut reprocher à la présente étude de n'aborder que la dimension prédictive des MPH, et d'intégrer des indicateurs de performances moins classiques. Ces faiblesses relèvent directement du contexte de l'étude et du but poursuivi par celle-ci.

Par ailleurs, comme il s'agit de la première étude abordant la PISA, d'autres tests et essais empiriques s'avéreront nécessaires pour juger de sa performance, appliquée à d'autres contextes et territoires d'analyses. Entre autres, il serait très intéressant d'intégrer la PISA dans un MPH spatio-temporel.

CONCLUSION

Dans un premier temps, cette étude conclut que les forces et facteurs influençant les prix hédoniques des emplacements se manifestent à la fois de manière continue et discrète, sur le territoire de Laval. Il appert donc que la segmentation optimale du territoire requiert d'inclure à la fois une segmentation *a priori*, délimitant les forces spatiales discrètes, et une mesure venant capter les phénomènes continus. À cet effet, nous concluons que la PISA, une approche novatrice introduite dans le présent mémoire, se veut une excellente candidate pour complémenter à une segmentation *a priori*, ne serait-ce que pour en peaufiner le découpage, ou encore mesurer certains effets omis par celle-ci. Une étude additionnelle est requise pour statuer si l'intervention d'un expert, ou encore la connaissance du territoire, est requise pour procéder à l'identification et la délimitation des forces discrètes affectant le territoire. La piste d'une approche centrée sur les données ne peut être formellement écartée, même si elle apparaît peu probable dans les secteurs où l'on retrouve des forces discrètes significatives et peu d'observations.

Dans un deuxième temps, une conclusion importante de ce mémoire est à l'effet que les unités de voisinage méritent bel et bien leur place prépondérante dans le processus d'évaluation municipale. À ce propos, le modèle EXPERT a généré la seconde meilleure performance prédictive dans la présente étude, tout juste derrière le modèle PISA. À l'opposé, nous émettons certaines réserves quant à l'utilisation des RGP en contexte d'évaluation municipale. Tout d'abord, il a été démontré que celles-ci subissent une importante perte de performance prédictive lorsqu'elles sont confrontées à l'estimation d'observations non utilisées pour calibrer les modèles. À cet effet, leur performance prédictive se classe à l'avant-dernier rang et se veut bien inférieure à celle des modèles EXPERT et PISA. Également, nous rapportons une autre problématique constatée avec les RGP, et rencontrée dans la littérature, à l'effet que certains coefficients locaux sont jugés irrationnels, c'est-à-dire non conformes aux

attentes en termes de valeurs ou de signes. À ce propos, il est important de noter qu'en contexte d'évaluation municipale, l'évaluateur doit fréquemment communiquer ses résultats d'évaluation et les vulgariser auprès de divers intervenants, qui ne sont pour la plupart ni experts en évaluation, ni en statistiques. Force est d'admettre qu'en pareil contexte, de tels coefficients contraires à la logique complexifient les échanges, et prédisposent à la critique. Or, la présente étude a abordé les RGP selon une approche relativement conventionnelle, c'est-à-dire avec un rayon adaptatif, une fonction de pondération géographique de type quasi-gaussienne fondée sur la distance euclidienne, et permettant à tous les coefficients de varier localement. Il est important de noter que les RGP permettent de nombreuses configurations alternatives que nous n'avons pas testées dans la présente étude : par exemple une fonction de pondération géographique basée sur la distance de route ou le temps de voyagement, ou une spécification permettant uniquement à certains coefficients de varier localement. Par conséquent, les résultats obtenus dans notre étude au niveau des RGP se doivent d'être interprétés avec discernement : s'il est vrai que les RGP n'ont pas généré les résultats escomptés dans notre étude, rien n'indique que la configuration que nous avons retenue est optimale. De plus amples études sont nécessaires pour tester davantage ces nombreuses configurations alternatives offertes par les RGP.

Pour conclure, la présente étude fait ressortir l'importance de la localisation dans la formation des prix immobiliers. Elle corrobore aussi bon nombre d'études attestant de la performance des MPH en évaluation immobilière. À cet effet, nous notons que le modèle PISA a généré, sur la base de l'ensemble des ventes, un taux d'erreur moyen de 5,44%, une erreur type de la régression de 0,069 et un R^2 de 93,99%. Qui plus est, celui-ci contrôle efficacement l'AS des résidus. Finalement, il ressort de notre étude que les MPH peuvent être améliorés par le recours à des algorithmes utilisés en data mining et, à plus forte raison, un brin de créativité.

RÉFÉRENCES BIBLIOGRAPHIQUES

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Dordrecht.

Bidanset P. E., Lombard J. R. (2014), Evaluating Spatial Model Accuracy in Mass Real Estate Appraisal: A Comparison of Geographically Weighted Regression and the Spatial Lag Model, *Cityscape: A Journal of Policy Development and Research*, Vol. *16*(3), 169–182.

Bitter, C., Mulligan, G. F., Dall'Erba S. (2007), Incorporating spatial variation in housing attribute prices: a comparison of geographically weighted regression and the spatial expansion method, *Journal of Geographical Systems*, Vol. *9*, 7-27.

Bourassa S. C., Hamelink F., Hoesli M., MacGregor B. D. (1999). Defining Housing Submarket, *Journal of Housing Economics*, Vol. *8*, 160-183.

Bourassa S. C., Hoesli M., Peng V. S. (2003) Do housing submarkets really matter? *Journal of Housing Economics*, Vol. *12*, 12-28.

Charlton M., Fotheringham A. S. (2009). Geographically Weighted Regression: White Paper, National Centre for Geocomputation, National University of Ireland Maynooth, https://www.geos.ed.ac.uk/~gisteac/fspat/gwr/gwr_arcgis/GWR_WhitePaper.pdf

Clapp J., Wang Y. (2006). Defining neighborhood boundaries: Are census tracts obsolete? *Journal of Urban Economics*, Vol. *59*(2), 259–284.

Des Rosiers F., Dubé J., Thériault M. (2011). Do peer effects shape property values? *Journal of Property Investment & Finance*, *29*(4/5), 510–528.

Dubin R. A., Sung C. H. (1990). Specification of Hedonic Regressions: Non-nested Tests on Measures of Neighborhood Quality, *Journal of Urban Economics*, Vol. *27*, 97-110.

Dubin R. A. (1998). Spatial Autocorrelation: A Primer. *Journal of housing economics*, Vol. *7*, 304-327.

Farber S., Yeates M. (2006). A Comparison of Localized Regression Models in a Hedonic House Price Context, *Canadian Journal of Regional Science*, XXIX: 3, 405-420.

Fotheringham A. S., Wong D. W. S. (1991). The modifiable areal unit problem in multivariate statistical analysis, *Environment and Planning*, Vol. *23*, 1025-1044.

Fotheringham A. S., Brunsdon C., Charlton M. (1998). Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment and Planning A*, Vol. *30*, 1905-1927.

Fotheringham A. S., Brunsdon C., Charlton M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley & Sons, Chichester.

Fotheringham A. S. (2009). "The Problem of Spatial Autocorrelation'' and Local Spatial Statistics. *Geographical Analysis*, Vol. *41*, 398–403.

Gao X., Asami Y., Chang-Jo C. (2006). An empirical evaluation of spatial regression models. *Computers & Geosciences*, Vol. *32*, 1040–1051.

Gehlke C. E., Biehl K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material, *Journal of the American Statistical Association*, Vol. *29*(185), 169–170.

Goodman A. C. (1981). Housing submarkets within urban areas: definitions and evidence. *Journal of Regional Science*, Vol. *21*(2), p. 175-185.

Goodman A. C., Thibodeau T. G. (1998). Housing Market Segmentation, *Journal of Housing Economics*, Vol. *7*, 121-143.

Gujarati D. N. (2004). *Basic Econometrics*. The McGraw−Hill Companies (4th edition), New York.

Helbich M., Wolfgang B., Hagenauer J., Leitner M. (2013). Data-Driven Regionalization of Housing Markets, *Annals of the Association of American Geographers*, *103*:4, 871-889.

Helbich M., Griffith D. A. (2016). Spatially varying coefficient models in real estate: Eigenvector spatial filtering and alternative approaches, *Computers, Environment and Urban Systems*, *57*, 1–11.

Hwang S., Thill J.-C. (2007). Using fuzzy clustering methods for delineating urban housing submarkets. *Proceedings of the 15th international symposium on advances in geographic information systems*. Article 14.

Hwang S., Thill J.-C. (2009). Delineating urban housing submarkets with fuzzy clustering. *Environment and Planning B: Planning and Design*, Vol. *36*, 865–882.

Kain J. F., Quigley J. M. (1970). Measuring the Value of Housing Quality, *Journal of the American Statistical Association*, Vol. *65*(330), 532-548.

Kestens Y., Thériault M., Des Rosiers F. (2006). Heterogeneity in hedonic modeling of house prices: Looking at buyers' households profiles. *Journal of Geographical Systems*, *8*(1), 61–96.

Le Gallo, J. (2002). Économétrie spatiale : L'autocorrélation spatiale dans les modèles de régression linéaire, *Économie et Prévision*, nº *155* 2002-4, 139-157.

Lockwood T. (2009). Delineation of Geospatial Residential Real Estate Submarket Boundaries. *Pacific Rim Property Research Journal*, Vol. *15*(1), 387-405.

Maclennan D. et Tu Y. (1996) Economic Perspectives on the Structure of Local Housing Systems. *Housing Studies*, Vol. *11*(3), 387-406.

Neyman J., Scott E. L. (1948). Consistent estimation from partially consistent observations. *Econometrica*, Vol. *16*(1), 1-32.

Páez A., Farber S. (2008). Moving Window Approaches for Hedonic Price Estimation: An Empirical Comparison of Modelling Techniques, *Urban Studies*, *45*(8): 1565-1581.

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, Vol. 82, pages 34 à 55.

Stock J., Watson M. (2012). *Principes d'économétrie* (Trad. par J. Trabelsi). Pearson (3ième édition).

Thériault M., Des Rosiers F. (2011). *Modeling Urban Dynamics*. John Wiley & Sons, Hoboken.

Tu Y., Sun H. et Yu S.-M. (2007). Spatial Autocorrelations and Urban Housing Market Segmentation. *The Journal of Real Estate Finance and Economics*, Vol. *34*, 385-406.

Voisin, M., Dubé J., Thériault M., Des Rosiers F. (2010). Les découpages administratifs sont-ils pertinents en analyse immobilière? Le cas de Québec, Cahiers de géographie du Québec, Vol. *54*(152), 249-274.

Wooldridge J. M. (2013). *Introductory Econometrics: A Modern Approach* (Fifth Edition). South-Western Cengage Learning, Mason.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*. Vol. *8*, 338-353.