LA MODÉLISATION STATISTIQUE EN ANALYSE ET ÉVALUATION IMMOBILIÈRES :

GUIDE MÉTHODOLOGIQUE

PAR

FRANÇOIS DES ROSIERS, PH.D.

GESTION URBAINE ET IMMOBILIÈRE

UNIVERSITÉ LAVAL

SEPTEMBRE 2001

1. PRÉFACE : CONTEXTE ET OBJECTIFS DU GUIDE MÉTHODOLOGIQUE

C'est au printemps de 1999 que se réunissaient pour la première fois les membres du *Groupe de travail sur les nouveaux développements relatifs à la méthode de comparaison*, mis sur pied par le Ministère des affaires municipales et de la Métropole (MAMM). Composé d'évaluateurs signataires de rôles, des représentants du Ministère et d'un expert du milieu universitaire, ce groupe de travail avait comme mandat le développement d'un nouvel outil visant à simplifier l'application de la méthode de comparaison et à en améliorer tant l'efficience que la performance. En vertu du système qui prévaut actuellement – et dont la philosophie repose toujours sur la méthode du coût de remplacement déprécié -, la quantité des données recueillies pour la confection des rôles résidentiels par la méthode de comparaison excède de beaucoup ce qui est nécessaire alors que la nature et la forme de ces données ne se prêtent pas toujours à un traitement efficace de l'information. L'objectif du groupe de travail était donc double : il s'agissait d'une part de restructurer en conséquence la fiche 2.6.1 du Volume 6 du *Manuel d'Évaluation foncière du Québec* (MÉFQ) et, d'autre part, de promouvoir le recours à l'approche statistique en évaluation de masse par la production d'un guide méthodologique qui en faciliterait l'usage par les utilisateurs éventuels. C'est précisément ce guide qui fait l'objet des chapitres qui suivent.

2. QUELQUES NOTIONS PRÉLIMINAIRES

2.1 L'ÉTABLISSEMENT DE LA VALEUR ET LE CONCEPT DE MARCHÉ

L'établissement de la valeur d'un bien est essentiellement issu du concept de marché : c'est le résultat de l'interaction de l'offre et de la demande, la première reflétant les conditions et contraintes de production de ce bien (coût des intrants, salaires, profit d'entreprise) et la seconde les caractéristiques des consommateurs (goûts et préférences, niveau des revenus, modes, etc.). Pour reprendre l'image d'Alfred Marshall dans son ouvrage célèbre "*Principles of Economics*"[1] , l'offre et la demande sont au marché ce que les deux lames sont au ciseau : c'est de la rencontre de ces deux forces que naît le prix du marché, lequel - il faut le souligner - n'est en fait *qu'une estimation de la valeur marchande.*

En effet, ce n'est que dans un marché de concurrence pure (marché atomistique) et parfaite (parfaite circulation de l'information) que le prix du marché pour un bien est identique à la valeur de ce bien. Or, en raison des caractéristiques propres aux biens immobiliers (immobilité physique, indivisibilité, hétérogénéité, information partielle), le marché pour ce type de produits s'écarte parfois substantiellement de ces conditions idéales et il devient alors nécessaire pour l'analyste d'estimer la valeur à partir des informations que fournit le marché, notamment les prix de transaction.

2.2 LA VALEUR MARCHANDE : UN CONCEPT STATISTIQUE

Il existe comme on le sait plusieurs définitions de la valeur d'un bien, lesquelles dépendent en particulier du but de l'évaluation. Ainsi, pour n'en citer que quelques-unes, la "valeur d'investissement" est la valeur subjective aux yeux d'un investisseur particulier; la "valeur de liquidation" est le prix payé dans une vente forcée; la "valeur d'expropriation" désigne la valeur d'usage (plutôt que la valeur d'échange) pour l'exproprié; la "valeur stabilisée", enfin, consiste en une valeur de long terme épurée des fluctuations excessives provenant de circonstances temporaires (crises, crash, pénuries, etc.). Pour l'évaluateur municipal chargé d'établir la valeur pour fins de taxation cependant, c'est le concept de valeur marchande qui s'impose comme critère de référence absolu.

On peut définir la "valeur marchande" d'un bien comme...

> *... le prix de transaction **le plus probable** auquel en arriveront un acheteur et un vendeur qui disposent de toute l'information nécessaire pour juger des caractéristiques du produit devant faire l'objet de la transaction (marché de concurrence parfaite) et qui agissent en toute indépendance et en toute liberté (transaction bona fide), sans que ni l'un ni l'autre ne soit en mesure d'exercer quelque forme de contrôle que ce soit sur le marché du bien en question (marché atomistique).*

Soulignons que cette définition de plus en plus largement retenue substitue l'expression "prix le plus probable" à celle de "prix le plus élevé" autrefois utilisée, ce qui ne fait que confirmer le caractère essentiellement statistique du concept de valeur marchande : en effet, alors que le prix de transaction demeure un événement isolé et circonstanciel, la valeur marchande réfère à un événement dont l'occurrence est *non pas certaine, mais probable* et qui, de ce fait, ne pourra se vérifier que sur un nombre relativement grand de cas.

1 A. Marshall, *Principles of Economics,* 8th. ed., Macmillan, London 1920.

2.3 VALEUR MARCHANDE, PRIX DE TRANSACTION ET COÛT DE PRODUCTION

De la section précédente on retiendra que prix de transaction et valeur marchande sont deux concepts différents : alors que le prix est le montant effectivement payé pour un bien, la valeur marchande est une estimation de ce qui aurait dû être payé dans les conditions d'un marché de concurrence pure et parfaite. A titre d'exemple, un acheteur mal averti pourra payer une propriété $300 000, qui n'en vaut en réalité sur le marché que $250 000; en ce sens, le prix « le plus élevé » n'est pas nécessairement « le plus probable », sauf s'il reflète un consensus de la part de l'ensemble des agents économiques quant au potentiel d'utilisation du bien concerné (concept d'usage optimal - "*highest and best use*").

Quant au "coût de remplacement" d'un bien, il représente un fait historique et peut, tout comme le prix de transaction, différer de la valeur marchande. L'écart entre la valeur marchande d'une propriété et son coût de remplacement peut provenir des déséquilibres inhérents à la nature même des marchés immobiliers : une forte demande pour un produit particulier qui n'est pas accompagnée d'un ajustement simultané de l'offre se traduira par une hausse de la valeur marchande (et des prix de transaction) pour ce produit qui excédera alors son coût de production, ce qui aura pour effet de générer à court terme des profits anormaux (sur-profits); l'inverse se produira en période de surproduction.

Si donc, à long terme, la concurrence, les ajustements de l'offre à la demande ainsi que la péréquation des taux de profits dans l'économie devraient faire en sorte que prix et coûts de production s'équilibrent (sans quoi les biens ne seraient plus produits), il n'en est pas nécessairement de même à court terme. Or, c'est toujours dans un contexte conjoncturel particulier qu'opère l'évaluateur, qui doit donc tenir compte des conditions marchandes régissant l'offre et la demande.

Cette distinction entre valeur marchande et coût de remplacement est cruciale pour le processus d'ajustement des comparables, que l'on peut à juste titre considérer comme le « talon d'Achille » de la méthode dite de parité.

2.4 LE PROBLÈME DE L'AJUSTEMENT DES COMPARABLES

L'évaluateur professionnel dispose essentiellement de trois techniques pour évaluer une propriété, soit la technique dite de "parité" (aussi appelée méthode des comparables), celle du "coût de remplacement déprécié" et celle du "revenu", cette dernière n'étant pertinente que dans le cas de propriétés à revenu. Alors que la méthode des comparables se veut une "preuve directe" de la valeur marchande dans la mesure où elle procède à partir de propriétés qui ont fait l'objet de transactions récentes et que la technique du revenu comporte également plusieurs éléments qui sont susceptibles de provenir directement du marché1 , la technique du coût déprécié reconstitue indirectement la valeur d'une propriété par addition des divers éléments qui la composent – i.e. les facteurs de production (sol, matériaux, main-d'oeuvre, capital financier) – , tout en l'ajustant pour tenir compte de la dépréciation physique, fonctionnelle et économique.

Cette dernière approche demeure, conceptuellement parlant, la plus critiquable : comme on l'a souligné plus haut en effet, si l'équivalence entre les coûts de production d'un actif immobilier et sa valeur marchande se vérifie sous certaines conditions (dans le cas d'un immeuble neuf par exemple), il n'en est pas toujours ainsi, la valeur marchande de l'actif pouvant, selon les cas, s'avérer inférieure, égale ou supérieure à son coût de remplacement. Or c'est précisément cette

1 Tels le «taux de rendement brut» (soit l'inverse du MRB), le «taux de rendement net» (soit l'inverse du MRN) et le «taux global d'actualisation» (R) dont le choix est crucial dans l'estimation de la valeur.

technique du coût de remplacement déprécié que l'on applique explicitement, dans l'approche paritaire, au processus d'ajustement des comparables , rendu pratiquement indispensable du fait de l'hétérogénéité des produits immobiliers. Par ailleurs, la perception qu'a l'évaluateur des marchés résidentiels et son opinion – si éclairée soit elle – de la valeur marginale des attributs d'une propriété introduisent inévitablement une part d'arbitraire dans l'établissement de la valeur marchande. Pour cette raison, le professionnel de l'évaluation cherchera dans la mesure du possible à définir son échantillon de référence (i.e. ses comparables) de manière qu'il soit le plus homogène possible et le plus représentatif de la propriété-sujet dans le but de minimiser le nombre des ajustements requis. Ce faisant, il risque toutefois de devoir réduire la taille de son échantillon en deçà du seuil minimal de fiabilité statistique et, partant, d'invalider les conclusions de son analyse.

C'est ici que le recours à l'approche statistique, et en particulier aux méthodes d'inférence statistique, revêt toute son importance.

3. LA MODÉLISATION STATISTIQUE EN ÉVALUATION IMMOBILIÈRE

PRÉSENTATION GÉNÉRALE

Cette publication présente une application méthodologique de l'utilisation de la méthode de comparaison, pour évaluer les terrains et certaines propriétés résidentielles. Elle est d'abord destinée aux évaluateurs signataires de rôles d'évaluation foncière, mais elle s'adresse également à leur personnel, ainsi qu'aux divers autres intervenants en la matière. Ce volume est constitué d'une première partie théorique comportant trois chapitres traitant respectivement des principes et concepts relatifs à la méthode de comparaison, de la méthodologie relative à l'application de la technique des prix de vente rajustés et de la méthodologie relative à la technique de modélisation statistique. Ce volume est également constitué de quatre autres parties pratiques dont deux sont à venir, soit la partie I, « Évaluation des terrains » et la partie III, « Évaluation des résidences multifamiliales », lesquelles parties seront ajoutées au fil des ans, en fonction des priorités de développement et de mise à jour du Manuel d'évaluation foncière du Québec (MÉFQ). Quant aux parties II et IV, elles traitent respectivement des résidences de trois logements et moins, de même que des copropriétés divises résidentielles. Elles contiennent les nouvelles fiches de propriété pour ce type de propriétés en plus des instructions administratives pour la présentation du dossier d'évaluation. Nous vous invitons à les utiliser et, au besoin, à nous formuler les commentaires ou suggestions relatifs à leur contenu. Au regard des besoins exprimés, des modifications pourraient être apportées lors des prochaines mises à jour du Manuel.

La réalisation de cet ouvrage a été possible grâce à la participation de plusieurs collaborateurs qui ont répondu avec empressement aux diverses consultations. Nous tenons à remercier particulièrement monsieur François Des Rosiers, Ph.D., professeur à l'Université Laval pour ses judicieux conseils : monsieur Des Rosiers est notamment l'auteur d'un document duquel s'inspire fortement le chapitre traitant de la modélisation statistique. La réalisation de cette partie du volume 6 n'aurait pas été possible sans son excellente collaboration. De nombreux autres intervenants méritent également nos remerciements, car si cette mise à jour a pu voir le jour, c'est grâce à l'implication et à la détermination de plusieurs personnes travaillant en évaluation foncière municipale. Nous tenons à remercier les nombreux collaborateurs externes (et bénévoles) pour leur acquiescement à nos sollicitations en matière de consultations, leur fourniture de renseignements et leur support technique. Ces remerciements s'adressent également aux responsables des firmes et organismes d'où proviennent ces personnes, lesquels ont dégagé les ressources nécessaires à la mise en oeuvre de cette fructueuse collaboration. Il s'agit de :

L'Association des évaluateurs municipaux du Québec;
L'Ordre des évaluateurs agréés du Québec;
Le Service d'évaluation de la Communauté urbaine de l'Outaouais;
Le Service d'évaluation de la Communauté urbaine de Montréal;
Le Service d'évaluation de la Communauté urbaine de Québec;
Le Service d'évaluation de la Ville de Lévis;
La firme Évimbec ltée;
La firme des estimateurs professionnels Leroux, Beaudry, Picard et associés.

TABLE DES MATIÈRES

3. LA MODÉLISATION STATISTIQUE EN ÉVALUATION IMMOBILIÈRE

3.1 INTRODUCTION

Ce chapitre propose *une démarche en dix étapes* pour modéliser statistiquement divers segments du marché immobilier. Il est construit autour d'une application empirique de l'analyse de régression portant sur le marché résidentiel unifamilial d'un cas réel du territoire québécois. Le modèle est présenté au point 3.5 et la démarche est couverte de façon systématique aux points 3.6 à 3.15.

Les concepts fondamentaux relatifs à l'application de la statistique en évaluation foncière, et plus particulièrement de l'analyse de régression, font, au préalable, l'objet des points 3.2, 3.3 et 3.4. Les divers aspects et problèmes soulevés lors du processus de modélisation sont abordés progressivement au cours de la démarche, au fur et à mesure des traitements réalisés à l'aide d'un logiciel statistique (SPSS)[1]. Les tableaux de résultats et les graphiques intégrés au texte en sont, par ailleurs, directement tirés, de façon à ce que le lecteur puisse se familiariser avec la démarche. Quant aux commandes du logiciel SPSS, elles sont présentées en annexe de ce chapitre afin de permettre à tout utilisateur éventuel de refaire la même démarche.

Il est à espérer que la démarche empirique, adoptée ici, permettra d'en démystifier l'usage et de démontrer au lecteur que les outils de statistique, tant descriptifs qu'analytiques, présentent un énorme potentiel dans l'exécution de ses tâches quotidiennes. Si, aux termes d'une lecture méthodique du document, il en ressort convaincu que ces concepts « hermétiques » peuvent être assimilés et correctement manipulés, moyennant un investissement intellectuel et financier relativement modeste, l'objectif aura été atteint. Encore faut-il évidemment qu'il accepte de s'y investir et de consacrer quelques jours à cet exercice. Toutefois, le principal défi qu'aura à relever l'analyste ne tient pas qu'aux dimensions techniques de l'approche statistique et de la modélisation statistique; il relève plutôt de la nouvelle philosophie que requiert cette approche eu égard à l'étude des marchés urbains et immobiliers, laquelle permet une meilleure vision d'ensemble que celle prévalant à l'application des méthodes traditionnelles. Comme il est permis de le constater, il est tout à fait possible d'atteindre d'excellentes performances à l'aide d'un nombre limité de variables, et ce, sans qu'il soit nécessaire de recourir à une structure fonctionnelle complexe. À cet égard, les résultats obtenus, avec le marché résidentiel unifamilial de la ville sous étude, sont assez typiques des performances déjà obtenues à l'aide d'échantillons similaires.

En conséquence, ce chapitre constitue essentiellement *un outil de vulgarisation* et non un ouvrage statistique spécialisé. Les expressions mathématiques (formules et équations) sont réduites au strict minimum et se limitent à celles qui sont incontournables, pour exposer les concepts statistiques indispensables à la compréhension du sujet de la façon la plus claire et la plus concise possible. Ce faisant, l'objectif est de faire en sorte que ce chapitre demeure très convivial d'utilisation et devienne une référence pour tout utilisateur de l'approche statistique en analyse et en évaluation foncière municipale, quels que soient ses antécédents en mathématiques et en statistique.

Il existe sur le marché un grand nombre d'ouvrages et de manuels de statistique de qualité permettant d'approfondir l'étude et la compréhension des concepts couverts dans ce chapitre.

1 À noter qu'il existe un excellent guide d'utilisation du logiciel SPSS, produit par Fernando Ouellet et Gérald Baillargeon et intitulé « *Traitement de données avec SPSS pour Windows, édition 8.0* ». Disponible aux éditions SMG.

Bien qu'il ne soit pas possible de fournir une liste exhaustive de ces ouvrages, voici quelques suggestions :

***a)* Ouvrages généraux :**

- Kazmier, Léonard J., *Statistiques de la gestion – Théorie et problèmes*, Série Schaum, Éd. McGraw-Hill, 1982, 374 pages.

- Sanders, Donald H., A. F. Murph et R. J. Eng, *Les statistiques : une approche nouvelle*, Éd. McGraw-Hill, 1984, 453 pages.

- Martel, Jean-M. et Raymond Nadeau, *Statistique en gestion et en économie*, Éd. Gaétan Morin, 1988, 621 pages.

***b)* Ouvrages spécialisés :**

- Neter, J., W. Wasserman and M. H. Kutner, *Applied Linear Statistical Models*, 2nd ed., R. D. Irwin, Homewood, IL, 1985, 555 pages.

- Kennedy, Peter, *A Guide to Econometrics*, The MIT Press, Cambridge, Mass., 1979, 175 pages.

Il existe également, au sein de certains ouvrages spécialisés en évaluation immobilière, des chapitres consacrés à l'analyse statistique. Mentionnons, en particulier :

- International Association of Assessing Officers (IAAO), *Property Appraisal and Assessment Administration*, Ed. Joseph K. Eckert, Chicago, Ill., 1990, chap. 14 et 15, p. 315-398.

- Desjardins, Jean-Guy, *Traité de l'évaluation foncière*, Éd. Wilson & Lafleur ltée (épuisé), Montréal, 1992, chap. 13 à 16, p. 437-513.

- Achour, Dominique, *Évaluation immobilière – Principes, concepts et pratiques*, Éd. Fisher Presses inc./Agence d'Arc, 1992, Appendice 1, p. 245-270.

3.2 APPROCHE STATISTIQUE APPLIQUÉE À L'ÉVALUATION DE MASSE

L'objectif premier du professionnel de l'évaluation foncière municipale demeure l'établissement de la valeur réelle d'une unité d'évaluation, en tenant compte des caractéristiques de cette unité. Or, la valeur réelle découlant, le plus souvent, d'un concept statistique, c'est par le recours aux méthodes d'analyse statistique qu'il est permis, avec le plus de fiabilité possible, d'atteindre le résultat désiré. L'utilisation de l'outil statistique peut évidemment se faire à divers niveaux : il est possible, par exemple, d'y recourir pour décrire une distribution de fréquence et d'en analyser les particularités ou anomalies (statistique descriptive); il est possible, également, de pousser plus loin l'usage de cet outil et l'utiliser à des fins prédictives ou pour expliquer un phénomène particulier (statistique analytique).

Le recours à l'instrument statistique implique que soient respectées un certain nombre d'hypothèses, en plus d'imposer certaines contraintes à l'utilisateur. Ce dernier se voit cependant doté d'un outil d'analyse extrêmement puissant et versatile, dont les avantages l'emportent indiscutablement sur les inconvénients. Par ailleurs, le développement de la technologie micro-informatique et la profusion des logiciels de traitement statistique rendent infiniment plus aisé et plus alléchant le recours à cette approche, dorénavant à la portée de tout analyste.

Il importe de ne pas sous-estimer le rôle que la statistique est appelée à jouer dans le champ de l'évaluation immobilière, et en particulier de l'évaluation résidentielle de masse qui constitue en la matière un domaine de prédilection : en effet, l'élaboration de banques de données de plus en plus sophistiquées, la recherche d'une meilleure équité fiscale au niveau local et la mise au point, par plusieurs organismes municipaux et régionaux, de systèmes intégrés d'information à références spatiales sont autant de facteurs militant en faveur d'une utilisation plus extensive de la statistique.

3.3 DÉFINITION DE L'APPROCHE HÉDONIQUE

Le logement est un bien *hétérogène* qui devient, lors de l'acquisition d'un immeuble, un « panier résidentiel » que se procure un ménage. Il se compose d'une certaine combinaison d'attributs physiques, socio-économiques, localisateurs et environnementaux faisant la spécificité de cet immeuble. La question qui se pose alors est double :

1. Comment mesurer la contribution respective de chaque caractéristique à la valeur réelle de l'immeuble?

2. Comment reconstituer la valeur réelle d'une unité d'évaluation à partir de ses attributs?

C'est précisément ce que permet de faire l'approche par modélisation statistique qui s'appuie sur plusieurs méthodes et théories ayant, depuis longtemps, acquis leurs lettres de noblesse, dont le calcul différentiel, la théorie des probabilités et la théorie micro-économique. La fonction de prix modélisé décrit en fait la relation entre le prix d'équilibre d'un bien hétérogène et ses caractéristiques. En résumé, l'approche modélisée permet d'isoler par analyse économétrique, plus précisément, par le recours à l'analyse de régression linéaire multiple, la valeur contributive de chaque attribut résidentiel, exprimée en termes absolus (forme linéaire) ou relatifs (forme multiplicative).

L'approche par modélisation statistique est mieux connue sous le terme « hédonique », en vertu de laquelle la contribution marginale (ou prix implicite) d'un attribut résidentiel est mesurée à partir de l'utilité (ou degré de satisfaction) qu'il procure au ménage. Cette utilité se reflète dans le prix du marché de l'actif immobilier. Bien que développée aux Etats-Unis dès le début des années 1940, cette approche a pris son envol avec les travaux de Rosen (1974)[1]. Le développement fulgurant des technologies de l'information a propulsé cette approche au premier rang des outils d'analyse de type CAMA (Computer-Assisted Mass Appraisal systems). Aujourd'hui utilisée un peu partout dans le monde (États-Unis, Royaume-Uni, Scandinavie, Australie, Asie, etc.), pour l'évaluation de masse et la mesure des externalités urbaines, cette approche est considérée comme une méthode très fiable pour l'étude de la formation des valeurs immobilières (en présence de marché). Elle présente, en effet, des avantages indéniables sur les approches traditionnelles utilisées en évaluation immobilière et sur certaines approches contemporaines non statistiques :

1. Cette méthode repose sur les propriétés du calcul différentiel qui, par le biais de la régression linéaire multiple, permet d'expliquer *les différences de prix observées sur un grand nombre d'immeubles vendus par leurs différences de caractéristiques* et d'isoler ainsi l'impact sur la valeur de chacun des attributs résidentiels *en tenant compte des « influences croisées »* entre les diverses variables de l'équation de régression.

2. L'approche statistique, faisant appel à la théorie des probabilités et donc à la loi des grands nombres, permet de valider chacun des résultats obtenus à l'aide *d'une batterie de tests* portant à la fois sur les performances globales (explicatives et prédictives) du modèle et sur le prix « implicite » relatif à chaque attribut résidentiel. Il est donc permis d'établir non seulement l'ampleur d'une contribution marginale positive ou négative, mais également son degré de fiabilité.

[1] Rosen, S. (1974) Hedonic prices and implicit markets: product differentiation in pure competition, *Journal of Political Economy*, **82**, 34-55.

3. Enfin, elle assure à l'analyste *une lecture directe et nuancée* des marchés immobiliers permettant de générer des indicateurs objectifs de la valeur, sans biais, distorsion ou ingérence externe, améliorant de ce fait la cohérence et l'équité des rôles d'évaluation réalisés à l'aide de cette méthode.

En dépit des complexités théoriques et méthodologiques de cette approche, il importe de rappeler ici qu'elle reproduit en réalité la psychologie des agents économiques qui, implicitement et à leur propre insu, font mentalement le même exercice. En effet, lors de la recherche d'une propriété, l'acheteur potentiel est confronté à différentes options d'achat impliquant des types de propriétés et des choix de localisation possiblement fort variés, les prix demandés par les vendeurs variant en conséquence. Après un certain nombre de visites, il est à même de se faire une idée assez précise de la valeur marchande de ces différentes options, sans toutefois pouvoir mettre un prix sur chacune des composantes du produit. Il solutionne toutefois cette équation par différenciation, ce qui lui permet de se prononcer sur le montant qu'il est prêt à payer pour un ensemble de caractéristiques données. Le prix offert par l'acheteur tient alors compte de ses besoins familiaux, de ses préférences personnelles et de ses contraintes budgétaires.

Quant à la valeur réelle d'une propriété dans un contexte donné, elle est le fruit de la négociation, entre acheteurs et vendeurs, et reflète *la dominante des valeurs subjectives* ainsi déterminées individuellement et soumises aux conditions locales de l'offre de l'immeuble (rareté relative). Les profils démographiques et socio-économiques, de même que les changements dans les préférences résidentielles des ménages exercent donc une influence déterminante sur les valeurs réelles.

À la suite de ce qui précède, il est facile de comprendre toute l'importance qu'il faut accorder à l'analyse des marchés et sous-marchés immobiliers, et au comportement des agents économiques (vendeurs et acheteurs). L'approche par modélisation statistique s'avère l'outil idéal pour y parvenir, puisqu'elle permet de « sonder les cœurs » des ménages à partir des ressources financières dont ceux-ci sont disposés à consacrer à la consommation des divers attributs d'un immeuble.

L'approche par modélisation statistique fait partie, avec les systèmes d'information géographique, des outils contemporains avec lesquels tout analyste immobilier doit se familiariser[1]. Bien sûr, cette approche n'est pertinente qu'en présence de marché, donc lorsque le nombre de transactions est suffisant. Ce problème ne se pose pas en région urbaine, du moins pour la catégorie des résidences unifamiliales. L'abondance et la qualité des bases de données transactionnelles, qui sont à la disposition de l'évaluateur municipal, militent en faveur d'une réorientation professionnelle vers une analyse plus globale et plus systématique des marchés immobiliers, à l'aide d'outils appropriés. Il en va de l'efficience et de l'équité du processus d'évaluation.

1 L'ouvrage de référence américain par excellence en évaluation immobilière, Property Appraisal and Assessment Administration, publié par l'International Association of Assessing Officers (IAAO, 1990), en fait largement mention (voir notamment chap. 15, p. 367-389). Au Québec, l'ouvrage de Desjardins (1992), Traité de l'évaluation foncière, y consacre un chapitre entier (chap. 15, p. 475-513).

3.4 RÉGRESSION LINÉAIRE : QUELQUES PRINCIPES FONDAMENTAUX

En statistique, la régression linéaire est une mesure de la relation existant entre deux ou plusieurs variables. Cette relation peut également se définir par une interdépendance des éléments observés sur le marché.

Essentiellement, la technique de régression linéaire repose sur l'hypothèse qu'à défaut de pouvoir établir une relation fonctionnelle exacte entre une variable dépendante (aussi appelée « endogène » ou « expliquée ») et une ou plusieurs variables indépendantes (aussi appelées « exogènes » ou « explicatives ») reflétant des phénomènes observables, il est possible d'estimer une telle relation statistiquement, à partir d'une série d'observations, grâce à l'introduction d'un terme lié au hasard (terme stochastique), permettant de prendre en considération les erreurs d'estimation. C'est ce terme qui fait la différence entre une relation mathématique *exacte* et l'estimation *probabiliste* des paramètres de l'équation de régression.

Dans le cas de la régression simple, dans laquelle se retrouve une seule variable indépendante, la relation peut s'écrire de la façon suivante :

$Y = B_0 + B_1X + e$, où :

- Y = variable dépendante
- X = variable indépendante
- B_0 = ordonnée à l'origine (constante)
- B_1 = coefficient de régression (paramètre estimé)
- e = terme stochastique (erreurs d'estimation)

3.4.1 MÉTHODE DES « MOINDRES CARRÉS ORDINAIRES » ET RÉGRESSION LINÉAIRE SIMPLE

Suivant le principe exposé plus haut, la relation linéaire entre « X » et « Y » peut être représentée par une droite traversant un nuage de points (c.-à-d. les valeurs observées), la distance entre ces points et la droite de régression constituant les erreurs (aussi appelés « résidus »). L'objectif est évidemment d'estimer les paramètres de cette droite de manière à minimiser les erreurs; la méthode généralement utilisée pour y parvenir est celle dite des « moindres carrés ordinaires » (*Ordinary Least-Squares ou OLS*), laquelle consiste à définir une droite qui soit telle que la somme des carrés des erreurs soit minimale. Puisque les erreurs (e_i) sont définies comme étant la différence entre les valeurs observées de la variable dépendante (Y_i) et les valeurs estimées par l'équation de régression (Y_i), cette somme des carrés des erreurs peut se poser comme suit :

$$\mathbf{e_i = Y_i - Y_i} \quad \text{et} \quad \mathbf{?} \quad \mathbf{e_i^2 = Minimum}$$

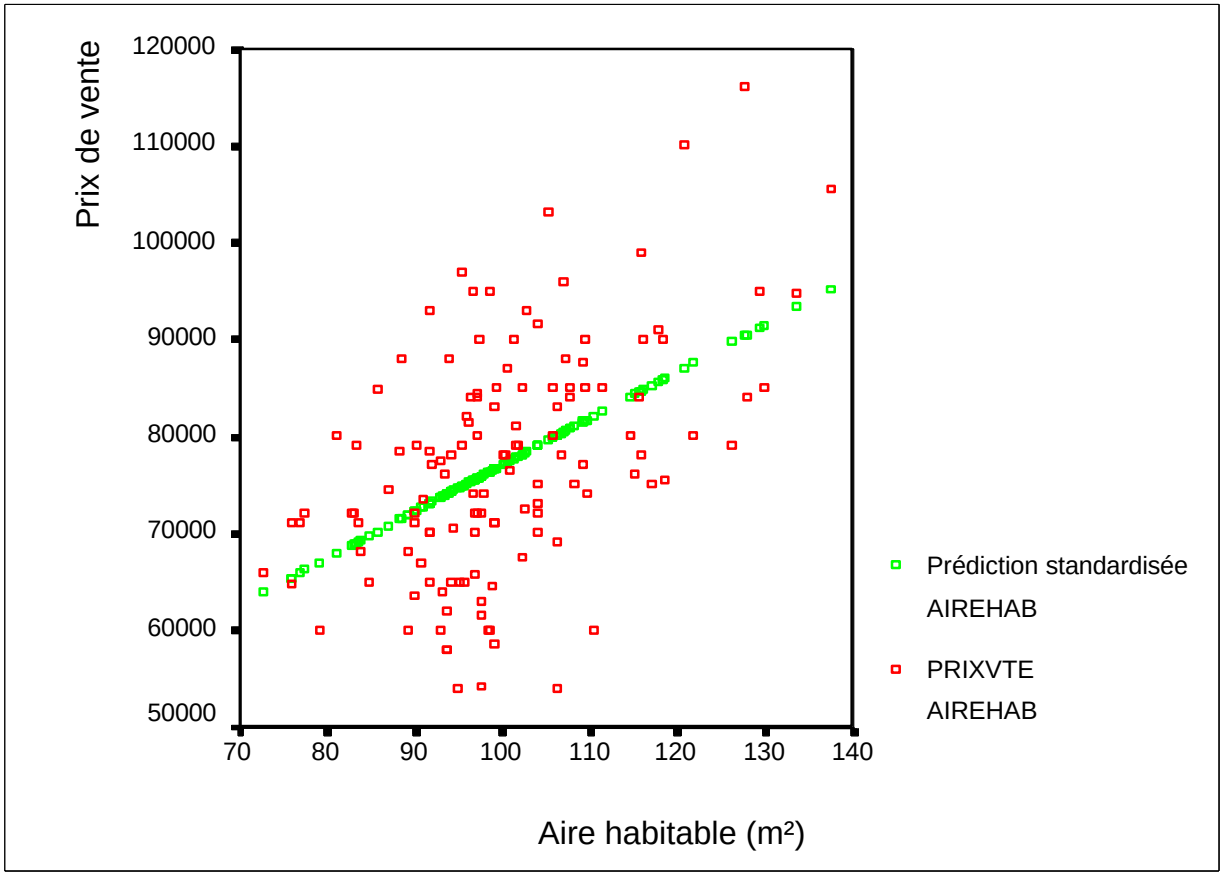
La droite de régression obtenue par la méthode des moindres carrés ordinaires appliquée à deux variables peut être représentée par un graphique à deux dimensions dans lequel l'axe des ordonnées supporte les valeurs de « Y » alors que les valeurs de « X » sont en abscisse.

3.4.2 EXEMPLE D'APPLICATION

Le *graphique 3.1* suivant reproduit la relation linéaire entre une série de prix de vente de résidences unifamiliales (Y_i) et l'aire habitable de ces immeubles (X_i), exprimée en mètres carrés. Dans ce graphique, le nuage de points, constitué ici de carrés vides, représente l'ensemble des 132 observations qui seront ultérieurement utilisées pour monter le modèle d'évaluation. La droite de régression simple apparaît en gris et se compose des valeurs prédites par l'équation de régression.

Bien que l'aire habitable soit une variable déterminante dans la formation des prix résidentiels, elle n'explique qu'une partie seulement des variations de prix observés. L'ordonnée à l'origine, qui constitue le terme constant, capte ainsi de façon globale l'influence des autres caractéristiques prises en compte par le modèle, mais non intégrées à l'équation comme variables explicatives. La portion des variations de prix ne pouvant être expliquée se retrouve dans les erreurs du modèle de régression; ces derniers se mesurent par la distance verticale entre chacun des points du nuage et la droite de régression.

Graphique 3.1 - Représentation graphique de la régression linéaire simple



Le coefficient de régression B_1, constituant en l'occurrence le prix « implicite » relatif à la variable « aire habitable », c.-à-d. la contribution à la valeur de chaque m^2 additionnel d'aire habitable, représente aussi *la pente* de la droite de régression.

3.4.3 RÉGRESSION LINÉAIRE MULTIPLE

La régression linéaire multiple n'est que l'extension du modèle à deux variables exposé précédemment. Elle permet d'étudier la relation entre une variable dépendante (Y) et un ensemble de plusieurs variables indépendantes (les X_i) considérées simultanément et constituant autant de dimensions (ou vecteurs) d'analyse. L'équation de base prend alors la forme suivante :

$$\mathbf{Y = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + ... + B_k X_k,}$$ où :

$X_1, X_2, X_3, ..., X_k$ = variables indépendantes

B_0 = ordonnée à l'origine

$B_1, B_2, B_3, ..., B_k$ = coefficients de régression non standardisés1

La régression linéaire multiple peut être utilisée dans un but explicatif, lorsque l'objectif est de décomposer un phénomène en ses éléments constitutifs et d'en déterminer la contribution respective, ou à des fins essentiellement prédictives, dans quel cas l'analyste s'intéressera surtout à la capacité des variables indépendantes, prises comme un tout, de déterminer « Y ». Il est à noter qu'**un modèle peut s'avérer relativement performant en tant qu'instrument de prédiction sans l'être nécessairement au plan explicatif**. Des tests statistiques permettant d'établir cette distinction seront présentés plus loin.

Il importe également de souligner que l'analyse de régression multiple suppose l'existence d'une relation de cause à effet entre, d'une part, la variable dépendante et, d'autre part, chacune des variables indépendantes. Toutefois, ce lien de causalité doit être établi *a priori* par l'analyste sur la base de quelque fondement théorique et n'est en rien garanti par la technique elle-même, laquelle ne fait qu'analyser dans quelle mesure les patterns de fluctuation des variables du modèle sont corrélés les uns aux autres.

Ceci dit, et dans la mesure où le modèle est bien spécifié (c.-à-d. lorsque l'introduction des variables explicatives est fondée et qu'aucune variable importante n'est omise), la méthode de régression multiple peut s'avérer d'une très grande utilité dans l'explication d'un phénomène. En particulier, elle possède sur la régression simple le grand avantage de permettre d'isoler la contribution relative d'une variable indépendante donnée en contrôlant l'influence exercée par les autres variables du modèle, lesquelles sont maintenues constantes grâce aux propriétés du calcul différentiel. Ainsi, expliquer la valeur marchande d'une unité d'évaluation par le secteur géographique (X_1) auquel elle appartient, par son âge (X_2) et par son état (X_3) implique que soient éliminées, d'abord, les influences « croisées » qu'exercent X_1 sur X_2, X_2 sur X_3 et X_1 sur X_3 de façon à établir l'apport explicatif marginal propre à chacune de ces dimensions, « *toutes choses étant égales par ailleurs ».* C'est précisément ce que peut accomplir la régression multiple.

---

[1] Lorsque les variables « Y » et « X_i » du modèle sont standardisées, c.-à-d. lorsqu'elles sont d'écarts types unitaires, la régression produit des coefficients dits « standardisés », dénotés par le symbole β*i.* Ces coefficients constituent en fait des poids relatifs qui, bien que ne pouvant être utilisés pour prédire « Y », permettent, cependant, de mieux faire ressortir la contribution respective de chacune des variables dans l'explication globale du phénomène étudié.

3.4.4 HYPOTHÈSES SOUS-JACENTES À LA MÉTHODE DE RÉGRESSION LINÉAIRE MULTIPLE

La méthode de régression linéaire multiple reposant sur la théorie statistique et la notion de probabilité, son application demeure sujette au respect d'un certain nombre d'hypothèses reliées *aux propriétés de la distribution normale*, laquelle domine l'économétrie traditionnelle et conditionne la validité des résultats obtenus avec l'approche par modélisation statistique. La distribution normale de probabilité, se traduisant graphiquement par la fameuse « courbe en cloche », est l'un des concepts les plus fondamentaux en statistique, du fait qu'un grand nombre de méthodes et d'applications statistiques reposent sur cette loi. Cela tient en fait aux raisons suivantes :

- les mesures obtenues dans de nombreux processus aléatoires peuvent être considérées comme obéissant à une loi normale;

- sous certaines conditions assez fréquemment rencontrées, une distribution normale de probabilité peut être utilisée comme une bonne approximation d'une autre distribution statistique (p. ex. : distribution binomiale ou distribution de Poisson);

- la distribution de certaines statistiques, telle la moyenne d'un échantillon, obéit souvent à une loi normale de probabilité, quelle que soit la distribution de la population sur laquelle est prélevé l'échantillon.

La distribution normale présente le grand avantage de n'être tributaire que de deux paramètres, soit la moyenne d'une distribution (μ) et sa variance (s^2). Elle possède, par ailleurs, les propriétés suivantes :

- la loi normale étant une loi de probabilités, l'aire sous la courbe et l'axe horizontal est de « 1 » (ou 100 %);

- la courbe normale est symétrique par rapport à la moyenne, laquelle divise donc l'aire en deux portions égales;

- puisque la courbe est symétrique, il est possible de présumer que moyenne = médiane = mode;

- la loi normale étant entièrement définie par ses deux paramètres μ et s^2, on obtient donc une distribution normale différente pour chaque valeur de μ et de son écart type (s);

- l'axe des abscisses est une asymptote, c'est-à-dire, à mesure que la moyenne s'éloigne, la courbe se rapproche de l'axe horizontal sans, toutefois, jamais le toucher. Par ailleurs, 68,3 % des valeurs probables sont comprises à l'intérieur des bornes définies par $\mu \pm 1s$, alors que 95,4 % le sont à l'intérieur de l'intervalle $\mu \pm 2s$; donc, l'aire située sous la courbe au-delà de $\mu \pm 3s$ est négligeable (probabilité d'occurrence inférieure à 1 %).

Il est essentiel de préciser que, dans le cas d'une distribution s'éloignant sensiblement de la normale, les propriétés de la distribution normale ne s'appliquent pas et les outils statistiques qui en dérivent ne peuvent être utilisés sans en préciser les limites.

Pour l'évaluateur immobilier, le recours à la régression linéaire multiple requiert que certaines hypothèses particulières soient respectées. Elles sont résumées ci-après et assorties de commentaires appropriés[1].

3.4.4.1 Existence de données complètes et fiables

La qualité d'un modèle statistique étant intimement liée à la qualité des informations servant à le construire, l'évaluateur doit apporter un soin particulier à la confection de la base de données, à sa validation et, éventuellement, à sa mise à jour (voir le point 3.8.1 de ce chapitre).

3.4.4.2 Linéarité de la relation découlant de la modélisation statistique

En vertu de cette hypothèse, la contribution marginale d'un attribut à la valeur marchande de l'unité d'évaluation est constante, quelle que soit la valeur prise par la variable. Comme il est mentionné plus loin (voir le point 3.9.2 de ce chapitre), il est possible de transformer mathématiquement les variables pour tenir compte des relations non linéaires, fréquentes en évaluation foncière. Le recours à un modèle multiplicatif constitue également une solution au problème.

3.4.4.3 Caractère additif des termes de l'équation

Corollaire de l'hypothèse précédente, l'additivité des termes de l'équation de régression implique que la contribution marginale d'un attribut résidentiel donné n'est pas affectée par les autres variables du modèle. Dans les faits, il existe souvent une interaction entre la contribution des attributs quantitatifs d'une propriété (p. ex. : l'aire habitable) et celle découlant de ses éléments qualitatifs (p. ex. : son état ou sa localisation). Ici encore, la transformation des variables, en particulier le recours à des variables interactives, permet de pallier à cet inconvénient. Il est possible, également, de recourir à la *forme multiplicative* (voir le point 3.9.2 de ce chapitre) ou au *modèle hybride* [2].

3.4.4.4 Indépendance des variables explicatives

Le recours à la régression linéaire multiple requiert en principe que les variables explicatives du modèle soient indépendantes les unes des autres. Dans le cas contraire, il y a présence de *multicolinéarité*, un problème classique en évaluation de masse et qui tient à la nature même des données immobilières. Il importe ici de préciser que les effets pervers de la multicolinéarité (instabilité et incohérence des coefficients de régression; tests statistiques invalidés) sont fonction du degré de gravité du problème. Une certaine corrélation entre les variables explicatives est, à toutes fins pratiques, inévitable, comme en fait foi l'analyse des corrélations (voir le point 3.10 de ce chapitre), et sans conséquences fâcheuses. C'est la présence de

---

[1] Tiré de IAAO, *Property Appraisal and Assessment Administration*, Ed. J. K. Eckert, Chicago 1990, p. 385-388.

[2] En raison de sa complexité, le modèle hybride, combinant variables additives et multiplicatives, n'est pas couvert dans le présent chapitre. Le lecteur intéressé par cette approche pourra consulter l'ouvrage de l'IAAO, p. 388-389. Il pourra, également, consulter l'article suivant constituant un exemple d'application des modèles hybrides au marché locatif de la région de Québec :

Des Rosiers, François et Marius Thériault, « Implicit Prices of Rental Services: Modeling the Quebec Market », *Assessment Journal* , Vol. 1, no 4, juillet-août 1994, p. 47-60.

corrélations excessives (c.-à-d. supérieures à 80 %) caractérisant la multicolinéarité *imparfaite* qui risque de poser problème1. Quant à la multicolinéarité *parfaite*, impliquant l'existence d'une combinaison linéaire exacte entre deux ou plusieurs variables, elle ne peut être tolérée par la procédure de régression qui génère alors un message d'erreur. La multicolinéarité peut être mise en évidence par une analyse attentive des résultats de la régression et par l'application du test VIF (Variance Inflation Factor) dont il sera question plus loin (voir le point 3.11.1.1 de ce chapitre). La solution au problème passe par l'élimination de la ou des variables qui sont à la source de la colinéarité, par la substitution d'une variable interactive aux deux variables fortement corrélées ou par l'application de la procédure de régression par étape (Stepwise), éliminant automatiquement de l'équation toute caractéristique dont la contribution marginale à la valeur réelle n'est pas suffisamment significative.

3.4.4.5 Normalité dans la distribution des résidus

Les termes d'erreurs, ou résidus, du modèle de régression doivent, tout comme la variable dépendante, suivre une distribution normale. La violation de cette hypothèse a pour effet de fausser l'interprétation de l'erreur type de prévision, tant absolue que relative, puisque les intervalles de confiance ne correspondent plus aux propriétés de la courbe normale, telles que définies au point 3.4.4 de ce chapitre. De plus, la valeur *« F »* et les valeurs *« t »*, testant respectivement la performance globale du modèle et la fiabilité des paramètres de la régression, c.-à-d. les coefficients des variables explicatives, sont également affectées par le problème. Ce dernier tient parfois à des déficiences dans la base de données, mais plus généralement à une mauvaise spécification des variables ou de la forme fonctionnelle du modèle, qui doivent donc être revues en conséquence.

3.4.4.6 Constance dans la variance des résidus

En plus d'être distribués normalement, les termes d'erreurs doivent être caractérisés par une variance constante qui demeure donc indépendante du niveau de prix des immeubles vendus. En d'autres termes, les erreurs de prévisions commises sur les immeubles les plus dispendieux de l'échantillon ne doivent pas présenter une plus forte dispersion que celles caractérisant les unités bas de gamme. La violation de cette hypothèse, dénommée *hétéroscédasticité* (c.-à-d. variance non homogène), se traduit par un modèle de régression dont les estimations portant sur les unités de faible valeur sont moins fiables, parce qu'influencées par les immeubles plus luxueux. Plusieurs solutions existent pour contrer les effets de l'hétéroscédasticité, depuis la transformation de la variable dépendante, en la divisant par exemple par l'aire habitable, ou l'utilisation d'une forme fonctionnelle multiplicative, jusqu'au recours à la procédure WLS (*Weighted Least Squares*, ou Moindres carrés pondérés). Ceci dit, la prévention demeure la meilleure approche : d'une façon générale, une segmentation judicieuse des sous-marchés immobiliers minimise l'importance du problème. L'évaluateur veille ainsi à ce que l'étendue de la distribution des prix de vente ne soit pas excessive (voir le point 3.9.1.1 de ce chapitre), le segment modélisé demeurant relativement homogène.

3.4.4.7 Indépendance des termes d'erreurs

Les résidus du modèle de régression étant en principe liés au hasard, ils doivent être indépendants les uns des autres. Dans le cas contraire, ces résidus sont confrontés au

[1] C'est, par exemple, ce qui se produit lorsque, dans l'équation de régression, le nombre de pièces et le nombre de chambres sont intégrés simultanément; les coefficients qui en résultent présentent une amplitude aberrante et un signe contraire à la logique.

phénomène d'*autocorrélation des erreurs*. Dans les analyses en coupe instantanée (« cross-sectional analysis »), situation s'appliquant plus particulièrement ici, il arrive fréquemment qu'un phénomène, affectant l'activité économique d'une zone ou le comportement de ses résidants, ait des répercussions analogues sur les zones voisines : c'est l'autocorrélation *spatiale*. Omniprésente en analyse spatiale, elle ne fait que reproduire les forces structurant le territoire. Ainsi, un fort taux de criminalité, dans un secteur géographique donné, peut se traduire par une baisse des valeurs immobilières non seulement du secteur en question, mais également des secteurs adjacents. De la même façon, le niveau du revenu moyen des ménages d'un secteur d'analyse peut être relié à celui des ménages des secteurs immédiatement voisins, ce qui ne fait que traduire le souci d'une certaine homogénéité des clientèles relatives aux divers sous-marchés résidentiels.

L'autocorrélation spatiale est responsable d'une portion non négligeable du pouvoir explicatif des modèles multivariés d'évaluation, ce qui met en évidence le rôle prépondérant qu'exercent les influences de voisinage sur les valeurs résidentielles. La mauvaise spécification du modèle constitue une autre cause fréquente d'autocorrélation : ainsi, l'omission d'une variable explicative importante, elle-même autocorrélée, tout comme l'adoption d'une forme fonctionnelle incorrecte, pour décrire la relation entre la variable dépendante et les variables indépendantes, se répercuteront au niveau des erreurs, lesquelles deviennent, de ce fait, autocorrélées. En présence d'une forte autocorrélation, les coefficients de régression ne sont plus efficaces et les tests de signification statistique deviennent, par conséquent, invalides. C'est par l'analyse des résidus, qui sont des estimations des erreurs, qu'il est possible d'obtenir sur ces dernières quelques informations. Alors que l'autocorrélation *temporelle,* propre aux modèles en coupe transversale, peut être notamment détectée par le test Durbin-Watson (D-W)[1], la détection et le traitement de l'autocorrélation spatiale est plus complexe et implique le recours à des procédures des statistiques spatiales, telles que *l'analyse par surface de tendance* et le *« krigeage »* permettant de capter les phénomènes de structuration de l'espace et de les réinsérer dans l'analyse de régression sous forme de variables ou de facteurs explicatifs[2]. Il importe, toutefois, de signaler que le recours à de telles procédures, indispensable dans les modèles d'analyse urbaine élaborés au niveau régional, n'est pas nécessaire dans le cas des modèles d'évaluation se limitant, en général, à des territoires beaucoup plus restreints.

3.4.4.8 Représentativité de l'échantillon

La dernière condition d'application de la régression linéaire multiple requiert que l'échantillon de transactions, servant à construire le modèle d'évaluation, soit représentatif de l'univers des unités d'évaluation auquel il est destiné. Il n'est pas possible, par exemple, d'appliquer à des propriétés âgées les résultats d'un modèle construit exclusivement sur la base de propriétés neuves; ou à des propriétés de type « cottage », les paramètres obtenus à partir d'un échantillon composé uniquement de « bungalows ». Ce point met en relief l'importance de la procédure d'échantillonnage discutée plus loin. L'une des façons de s'assurer de la représentativité de l'échantillon est de comparer, à l'aide des statistiques descriptives (moyenne, écart type, distributions de fréquences, etc.), le profil des unités transigées le composant à celui des

[1] Le test D-W, se retrouvant sur tous les logiciels statistiques, n'est toutefois applicable qu'aux modèles temporels seulement et n'est donc pas pertinent dans le cas des modèles d'évaluation.

[2] Le lecteur intéressé à en savoir plus sur la question pourra se référer à l'article suivant :

Des Rosiers, François and Marius Thériault,« House Prices and Spatial Dependence : Towards an Integrated Procedure to Model Neighborhood Dynamics », *Working Paper # 1999-002*, Faculty of Business Administration, Laval University, December 1998.

immeubles non transigés de l'univers d'application du modèle statistique. Cela n'est évidemment possible que dans la mesure où il existe une description détaillée de chacune des unités d'évaluation composant le parc résidentiel visé par l'exercice de modélisation. Il sera possible alors de vérifier que le profil type des propriétés de l'échantillon reproduit fidèlement celui du parc cible. En outre, il est prudent de générer, outre l'échantillon principal, *un sous-échantillon contrôle* de transactions qui ne sont pas utilisées dans l'analyse de régression, mais serviront à valider le modèle final et à en vérifier la robustesse (voir le point 3.8.2 de ce chapitre). Le modèle est considéré comme étant robuste si les résultats qu'il produit, sur les immeubles de ce sous-échantillon, sont conformes à sa performance prédictive théorique.

3.4.5 FORCES ET LIMITES DE LA MÉTHODE DE RÉGRESSION LINÉAIRE MULTIPLE

L'analyse de régression constitue un outil d'analyse puissant et performant. En dépit de la complexité apparente des procédures statistiques qu'implique son utilisation, l'abondance des logiciels micro-informatiques existant sur le marché permet à quiconque veut s'en donner la peine de recourir à cette approche et d'obtenir, même en s'en tenant aux aspects les plus élémentaires de la technique, des résultats très satisfaisants. L'analyse de régression présente également certaines limites. Les avantages et les limites de la méthode se résument comme suit :

- elle peut établir dans quelle mesure les fluctuations d'une variable dépendante (Y) sont associées à celles d'une ou de plusieurs variables indépendantes (X_i) et décrire cette relation par le biais de l'équation de régression;

- appliquée à la problématique de l'évaluation foncière, elle permet d'établir la contribution marginale de chaque attribut résidentiel à la valeur réelle d'une unité d'évaluation et de reconstituer cette valeur à partir des caractéristiques de l'unité d'évaluation;

- du fait qu'elle repose sur la théorie des probabilités, elle bénéficie d'une batterie de tests statistiques permettant d'établir à la fois la performance globale d'un modèle, aux plans tant explicatif que prédictif, et la fiabilité des coefficients individuels relatifs à chaque variable;

- elle permet une lecture directe et nuancée des marchés et sous-marchés immobiliers assurant l'objectivité et la cohérence des valeurs obtenues ainsi que l'équité des rôles d'évaluation;

- l'approche par modélisation statistique, couramment utilisée aux États-Unis depuis plus de deux décennies pour la modélisation des valeurs immobilières et la confection des rôles d'évaluation, repose sur des bases théoriques et méthodologiques très solides et très bien documentées qui ont, depuis longtemps, acquis leurs lettres de noblesse;

- la versatilité de la méthode et la robustesse des résultats obtenus expliquent sa popularité croissante dans plusieurs provinces canadiennes, notamment en Saskatchewan et en Alberta, ainsi qu'en Europe et en Asie. L'énorme potentiel d'intégration entre l'approche statistique et les SIG (Systèmes d'information géographique) ne fait qu'en accentuer les avantages[1];

[1] À cet égard, voir : Roberto A. Figueroa, Modeling the Value of Location in Regina Using GIS and Spatial Autocorrelation Statistics, *Assessment Journal*, Nov.-Dec. 1999, p. 29-37.

- en contrepartie, les analyses de régression et de corrélation ne s'appliquent qu'à des échantillons de taille relativement grande;

- ces procédures d'inférence statistique n'impliquent *a priori* aucune relation de cause à effet entre les variables et demeurent essentiellement *des outils d'aide à la décision* ne pouvant servir de substitut au jugement de l'analyste, à sa bonne connaissance du marché local et à l'étude qualitative des faits;

- enfin, la qualité des résultats ne peut être meilleure que celle des données utilisées.

3.5 PROCÉDURE D'ÉLABORATION D'UN MODÈLE STATISTIQUE

L'élaboration d'un modèle statistique, pour fins d'évaluation foncière, requiert qu'une procédure rigoureuse soit suivie, de laquelle dépendra la qualité du produit final. Force est de constater que ce n'est pas toujours le cas et que les frustrations que peut, éventuellement, rencontrer l'analyste tiennent la plupart du temps au fait qu'il aura négligé certaines étapes de cette procédure. Il importe de bien connaître ces étapes, lesquelles sont présentées aux points 3.6 à 3.15 de ce chapitre.

3.5.1 ÉTAPES À SUIVRE

À la lumière de recherches sur le sujet, une démarche proposée en dix étapes réduit sensiblement le risque d'omissions majeures et d'erreurs méthodologiques[1]. Toutes ces étapes, résumées ci-après, doivent être respectées, qu'il s'agisse de la collecte de l'information, de la validation de la base de données, des diverses procédures de traitement ou de la validation finale du modèle.

Le tableau de la page suivante donne, dans l'ordre, ces dix étapes et dans les points suivants, celles-ci sont passées en revue pour les appliquer à l'analyse d'un marché résidentiel unifamilial. Quant à l'outil statistique utilisé tout au long de cette analyse, il s'agit du logiciel *SPSS*[2] *pour Windows, version 10.0*. Très complet et très versatile, ce logiciel présente également le grand avantage d'être particulièrement convivial. Il permet en outre d'importer facilement des données enregistrées sous format Excel ou DBF (Menu « ***File*** », commande « ***Database Capture*** », sous-commande «***New Query*** »). Enfin, il offre des possibilités graphiques et d'édition qui facilitent la visualisation des résultats de l'analyse et leur intégration au sein de documents textes. Il suffit en effet de double-cliquer sur les tableaux et les graphiques pour en modifier la présentation à volonté (choix de polices de caractères, format des tableaux, choix de couleurs, etc.). Le coût du logiciel SPSS varie selon le « kit » désiré : si la version commerciale standard du logiciel se détaille aux alentours de $2 000, la *version professionnelle* (*Graduate Pack – Business Version, v. 10.0)*, tout aussi complète et disponible dans les institutions d'enseignement supérieur, se vend $270. La *version étudiante* (*Student Version, v. 9.0)*, une version réduite qui n'accepte que 1 500 observations et 50 variables, s'avère largement suffisante pour répondre à la plupart des besoins en évaluation et peut être obtenue pour environ $100. Un résumé des commandes utiles à l'élaboration d'un modèle statistique avec ce logiciel est présenté en annexe de ce chapitre.

Il existe, bien sûr, un grand nombre de logiciels statistiques sur le marché pouvant effectuer les mêmes tâches que SPSS. Le logiciel StatView, fonctionnant sur les plates-formes PC et Apple, est l'un d'entre eux. L'apprenti statisticien doit s'assurer, au préalable, que les procédures utilisées dans le présent chapitre (notamment la procédure Stepwise), ainsi que les tests statistiques recommandés (notamment les VIF), sont bel et bien disponibles avant d'en faire l'acquisition.

---

[1] Il importe de préciser ici que seules les étapes *élémentaires* nécessaires à la construction de modèles d'évaluation sont prises en compte ici. Les modèles d'analyse urbaine requièrent une procédure plus sophistiquée faisant appel à diverses méthodes de statistique spatiale.

[2] À noter qu'il existe un excellent guide d'utilisation du logiciel SPSS, produit par Fernando Ouellet et Gérald Baillargeon et intitulé « *Traitement de données avec SPSS pour Windows, édition 8.0* ». Disponible aux éditions SMG.

ÉLABORATION D'UN MODÈLE STATISTIQUE : les étapes à suivre

1. **Définition des objectifs de la modélisation et approche analytique**

2. **Choix et description du secteur d'analyse et nature de l'échantillon**

   - Le type de segmentation
   - La représentation cartographique

3. **Collecte de l'information et définition des variables**

   - Le traitement de la base de données
   - La sélection d'un sous-échantillon pour valider ultérieurement le modèle
   - La définition opérationnelle des variables retenues

4. **Description et analyse de la base de données**

   - L'application des statistiques descriptives
   - La transformation mathématique de variables
   - L'application des statistiques descriptives après corrections et épuration

5. **Analyse de corrélation**

   - Le test de corrélation simple
   - Le test de fiabilité

6. **Analyse de régression**

   - La procédure de régression standard
   - La réduction du nombre de variables
   - La transformation mathématique des variables indépendantes
   - La transformation mathématique de la variable dépendante
   - La procédure de régression par étape

7. **Analyse des résidus**

   - L'identification des résidus délinquants
   - La représentation graphique des résidus

8. **Mise au point du modèle final**

9. **Validation du modèle final à l'aide du sous-échantillon retenu à l'étape 3**

10. **Production d'indications de la valeur**

3.6 ÉTAPE 1: DÉFINITION DES OBJECTIFS DE LA MODÉLISATION ET APPROCHE ANALYTIQUE

La modélisation statistique peut s'appliquer à plusieurs situations et viser divers objectifs. En particulier, elle est utilisée :

- pour *construire un modèle explicatif de la dynamique urbaine et immobilière de l'agglomération dans son ensemble*. Il est primordial alors de recourir à une base de données relativement volumineuse et utiliser un grand nombre de caractéristiques et attributs décrivant non seulement l'unité d'évaluation, mais également le voisinage, le profil socio-économique et démographique des ménages, les éléments d'accessibilité et de proximité aux services ainsi que les diverses externalités urbaines, tant positives que négatives;

- *pour mesurer la contribution spécifique d'une externalité donnée à la valeur marchande des unités d'évaluation* (p. ex. : l'impact de la proximité d'une autoroute, de lignes de transport hydroélectriques, d'une station de métro, etc.). Il est recommandé alors de mettre l'accent sur la mesure du phénomène étudié et la forme fonctionnelle permettant d'en capter la dynamique de façon optimale;

- *pour construire un modèle d'évaluation devant servir à la confection d'un rôle d'évaluation*. Sans négliger la performance explicative d'un tel modèle, il sera important de se préoccuper particulièrement de sa performance prédictive dont dépend l'équité du rôle d'évaluation. Ce genre de modèle, s'appliquant en général à un secteur bien délimité du territoire urbanisé ou à un type particulier de résidences, ne requiert qu'un échantillon restreint d'immeubles vendus et peut atteindre d'excellentes performances avec relativement peu de variables (une quinzaine ou moins).

Il est donc essentiel d'établir clairement, dans un premier temps, quels sont les objectifs recherchés par l'analyste dans son recours à l'approche statistique.

Les concepts relatifs à l'approche de modélisation statistique et à son utilisation en évaluation immobilière ont été exposés au point précédent. Il est toutefois important pour l'évaluateur, qui compte recourir à cette approche, d'en résumer brièvement la nature, les objectifs et les principes fondamentaux.

3.6.1 EXEMPLE D'APPLICATION

L'exemple d'application élaboré tout au long de ce chapitre vise à déterminer la valeur réelle, le premier juillet 1999, des résidences unifamiliales d'un étage de la municipalité en cause.

L'objectif de la modélisation est de construire un modèle d'évaluation devant servir à établir la valeur au rôle 2001-2002-2003 des résidences dont il est question, afin de les évaluer en fonction des mêmes critères d'évaluation.

Le modèle devra viser la même segmentation de marché (voir l'étape suivante) et atteindre une performance de prévision moyenne sous le seuil de 10 % (voir l'étape 9).

3.7 ÉTAPE 2 : CHOIX ET DESCRIPTION DU SECTEUR D'ANALYSE ET NATURE DE L'ÉCHANTILLON

À cette étape, l'évaluateur détermine le segment du marché local à modéliser. La réponse à cette question varie selon la taille de la municipalité et le degré de complexité de son parc. Le choix du secteur d'analyse dépend aussi de la tolérance de l'analyste face au degré de précision du modèle statistique. En effet, le niveau d'hétérogénéité de l'échantillon augmente en général avec l'ampleur du territoire englobé dans l'analyse. Aussi, délimiter un trop grand territoire regroupant en réalité plusieurs sous-marchés résidentiels, risque de détériorer la performance prédictive du modèle, à moins d'augmenter en conséquence le nombre de descripteurs (variables explicatives). Par contre, le choix d'un territoire trop restreint réduit l'univers d'application de l'instrument d'évaluation ainsi élaboré.

3.7.1 TYPE DE SEGMENTATION

La segmentation du marché local peut conduire à trois types de modèles, soit :

- *le modèle sectoriel*, où la segmentation s'effectue selon une base spatiale (p. ex. : par famille d'unités de voisinage);

- *le modèle typologique*, où la segmentation se fait selon le type de propriété (p. ex. : le sous-marché des « bungalows » ou des maisons unifamiliales détachées);

- *le modèle mixte*, combinant les deux approches précédentes (p. ex. : le sous-marché des « bungalows » dans une famille d'unités de voisinage).

Le type de segmentation est souvent dicté par des contraintes de marché, notamment l'importance du parc résidentiel et le volume des transactions. Comme le recours à la modélisation statistique requiert un échantillon d'au moins 50 à 60 ventes[1], une segmentation mixte, impliquant un découpage plus poussé du marché local, ne s'applique que là où le volume des transactions le permet. De plus, l'analyste aura avantage à bien maîtriser les étapes préalables à la méthode de comparaison, lesquelles sont présentées au point 1.5 du chapitre 1 de la partie générale de ce volume.

Par ailleurs, lors de la segmentation, les consignes suivantes doivent être respectées :

- le segment choisi doit respecter *une certaine homogénéité* : ainsi, l'inclusion simultanée des propriétés « haut de gamme » et des propriétés « bas de gamme » est évitée (p. ex. : prix variant de 50 000 $ à 350 000 $); bien qu'aucune norme formelle n'existe à ce sujet, il est important de tenter de limiter l'étendue de la distribution des prix à un maximum de 75 000 $. Le mélange trop prononcé des genres (p. ex. : maisons unifamiliales et unités en copropriété) est également à éviter. Toutes ces précautions ont pour effet de réduire les risques d'hétéroscédasticité. Dans tous les cas, l'échantillon doit être représentatif de l'univers visé par le modèle;

[1] En règle générale, au moins quatre observations, pour chaque variable indépendante, sont exigées. Ainsi, pour un modèle comportant quinze variables explicatives, un minimum de 60 transactions devra être recueilli (IAAO, op. cit., p. 343).

- le cas échéant, c'est-à-dire si le segment choisi comporte plusieurs sous-marchés spatiaux, *des sous-secteurs d'analyse,* permettant d'ajuster, en conséquence, les prix établis par le modèle, sont délimités;

- la délimitation des secteurs et sous-secteurs s'effectue sur la base des unités de voisinage (voir chapitre 4, volume 2 du MÉFQ). Les unités de voisinage sont regroupées en familles d'unités de voisinage en considérant les éléments dominants de *la trame urbaine* (axes de circulation majeurs, voie ferrée, cours d'eau, parc industriel, etc.), ou encore selon *des critères socio-économiques* (revenu moyen des ménages), *architecturaux* (correspondant aux phases de développement du parc résidentiel) ou *historiques* (anciens quartiers versus nouveaux quartiers);

- l'analyste, dans la mesure du possible, *identifie les externalités* tant positives (p. ex. : qualité de la vue) que négatives (p. ex. : nuisance visuelle, sonore ou environnementale) étant susceptibles d'affecter le prix des immeubles. Le cas échéant, ces externalités peuvent être intégrées au modèle de régression.

3.7.2 REPRÉSENTATION CARTOGRAPHIQUE

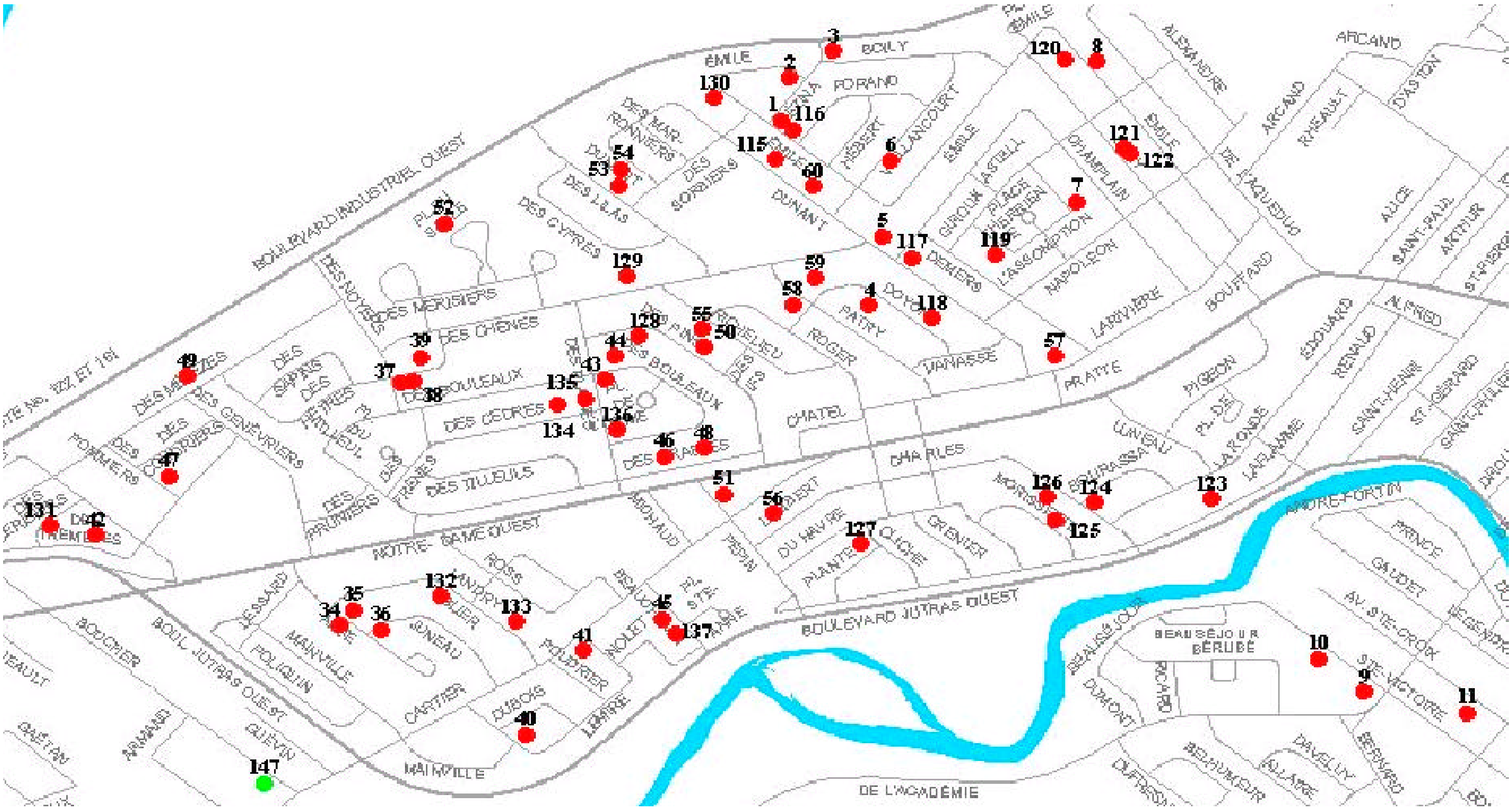
La segmentation s'accompagne *d'une représentation cartographique* du secteur et des sous-secteurs d'étude montrant la localisation des transactions retenues; le recours à un SIG (système d'information géographique) est évidemment souhaitable, bien que non indispensable à ce stade. L'utilisation de la matrice graphique est également un excellent moyen pour représenter les secteurs d'études, surtout lorsque cette matrice est supportée par un système géomatisé.

3.7.3 EXEMPLE D'APPLICATION

L'étude de cas, utilisée pour le présent chapitre, porte sur un *marché de « bungalows »*; il s'agit donc d'un modèle typologique. Toutes les transactions *bona fide,* répertoriées entre février 1997 et août 1999 dont la construction est de type « bungalow », soit 147, ont été retenues pour les fins de la modélisation. L'étendue de la distribution des prix de 62 000 $ (prix variant de 54 000 $ à 116 000 $) indique une certaine homogénéité des classes des immeubles du parc cible. De plus, les bâtiments sont de classe 4 ou 5, de type détaché, dont la date de construction apparente n'excède pas 30 ans alors que l'aire habitable se situe entre 70 et 140 m².

Trois sous-secteurs d'analyse sont délimités, soit les familles d'unités de voisinage *« A »* (79 ventes), *« B »* (50 ventes) et *« C »* (18 ventes). Les unités de voisinage sont regroupées en familles d'unités de voisinage sur la base d'anciens territoires fusionnés. L' hypothèse est à l'effet que les unités de voisinage comprises dans ces sous-secteurs évoluent avec les mêmes tendances.

La *Carte 3.1*, de la page suivante, reproduit une partie des secteurs d'analyse ainsi que la localisation et l'identification d'un bon nombre de transactions.


BOULEVARD INDUSTRIEL OUEST
BOULEVARD JUTRAS OUEST
BOUL. JUTRAS OUEST
NOTRE-DAME OUEST
DE L'ACADÉMIE
BEAUSÉJOUR
BÉRUBÉ
ANDRÉ-FORTIN
ARCAND
ALEXANDRE
DES CHÊNES
DES MERISIERS
DES CÈDRES
DES TILLEULS
DES BOULEAUX
DES CYPRÈS
DES LILAS
DES SORBIERS
CHARLES
CHATEL
PRATTE
VANASSE
DUNANT
BOILY
FORAND
MAINVILLE
CARTIER
DUBOIS
POLIQUIN
BOUCHER
LARIVIÈRE
BOUFFARD
NAPOLÉON
L'ASSOMPTION
CHAMPLAIN
DEMERS
PATRY
ROGER
GRENIER
PLANTE
DU HAVRE
STE-VICTOIRE
AV. STE-CROIX
GAUDET
PRINCE
LAFLAMME
LALONDE
PIGEON
RENAUD
EDOUARD
ALFRED
SAINT-HENRI
ST-GERARD
ALICE
ARTHUR
DASTON
RASSAULT
1
2
3
4
5
6
7
8
9
10
11
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
147

3.8 ÉTAPE 3 : COLLECTE DE L'INFORMATION ET DÉFINITION DES VARIABLES

L'information utilisée pour la modélisation statistique provient de 155 fiches de propriété, constituées des formulaires 2.6.1 C et 2.6.9 C, lesquels sont respectivement décrits au volume 2 du MÉFQ pour le premier formulaire et à la partie II du présent volume pour le deuxième formulaire.

Ces formulaires doivent être remplis selon les consignes prévues à cette fin. Les cases de ces formulaires comportent les données de base servant à définir les variables utilisées pour réaliser la modélisation.

3.8.1 TRAITEMENT DE LA BASE DE DONNÉES

Puisque la qualité des prédictions du modèle statistique reflète directement celle des informations servant à le construire (cf. le dicton « *Garbage in, Garbage out »*), il est primordial d'insister sur l'importance de cette étape. En premier lieu, *prendre soin de ne conserver que les ventes bona fide*, ce qui élimine d'office, notamment, les transactions entre parents, les ventes entre filiales, les ventes à 1 $, les cas de faillites, de reprises d'hypothèque, ainsi que les ventes de succession (voir le point 9.5.4, chapitre 9, volume 2 du MÉFQ). Enfin, il est possible d' imaginer certaines situations de marché où les reprises d'hypothèque ou les transferts professionnels sont la norme plutôt que l'exception (p. ex. : une ville minière en récession). Ces ventes sont alors conservées comme partie intégrante de la dynamique résidentielle locale.

Même si toutes les transactions retenues sont du type *bona fide*, plusieurs doivent être retirées de la base pour les fins de la modélisation *si elles ne rencontrent pas certains critères établis à l'étape antérieure*. Par exemple, l'étendue des prix peut être réduite en éliminant les extrêmes (propriétés bas de gamme et haut de gamme), en retirant les immeubles ayant une superficie habitable anormalement faible ou élevée, ou encore celles dont l'âge avancé en fait une catégorie à part (maison historique).

Enfin, *il est important de valider chaque inscription*, les erreurs de saisie étant fréquentes et pouvant s'avérer dommageables si elles portent sur des variables clés (prix de vente, aire habitable, âge chronologique ou apparent, superficie du terrain, etc.). Ces erreurs d'inscription sont éventuellement corrigées, puis l'observation réintégrée dans l'échantillon.

3.8.1.1 Exemple d'application

Dans le cas présent, huit observations sont retirées de la banque initiale dont 155 ventes étaient incluses à l'origine. Ces ventes ne rencontraient pas les critères de vente *bona fide.*

3.8.2 SÉLECTION D'UN SOUS-ÉCHANTILLON POUR VALIDER ULTÉRIEUREMENT LE MODÈLE

Bien que le recours à l'approche de modélisation statistique comporte en soi tous les outils statistiques permettant de valider le modèle obtenu, tant globalement qu'au niveau des contributions spécifiques à chaque attribut résidentiel, le test ultime consiste à appliquer le modèle à une série d'immeubles, dont le prix de vente et les caractéristiques sont connus, mais qui n'ont pas servi à son élaboration. Ces transactions « indépendantes » représentent environ 10 % de l'échantillon, après épuration, et sont extraites de la banque de départ en fonction d'une procédure d'échantillonnage pouvant varier selon les cas :

- *l'échantillonnage au hasard* consiste à sélectionner *« n»* cas parmi l'échantillon de taille *« N*[1] *»* suivant un processus aléatoire, c.-à-d. par l'intermédiaire d'un mécanisme probabiliste;

- *l'échantillonnage stratifié* consiste, dans un premier temps, à subdiviser l'univers d'analyse (l'ensemble des transactions retenues) en sous-groupes relativement homogènes (p. ex. : les sous-secteurs spatiaux), puis à procéder à un échantillonnage aléatoire dans chaque strate;

- en vertu de *l'échantillonnage systématique* enfin, les immeubles vendus à intervalles fixes sont sélectionnés (p. ex. : une toutes les dix observations).

3.8.2.1 Exemple d'application

Dans l'analyse suivante, quatorze ventes (sur un total de 147) sont isolées en vertu d'un échantillonnage systématique (ventes 10, 20, 30, …,140). Cette opération est facilitée par les fonctions de gestion des données du logiciel SPSS.

3.8.3 DÉFINITION OPÉRATIONNELLE DES VARIABLES

La dernière phase de l'étape 3 consiste à définir les variables qui seront utilisées pour la modélisation statistique. Bien que ces variables soient directement tirées des paramètres relatifs à l'approche comparative (voir les points 3.6 à 3.15 de ce chapitre), plusieurs doivent être adaptées pour les besoins de la modélisation statistique.

Les modèles statistiques ont en général recours à trois types de variables :

- *les variables métriques (M)* sont des variables quantitatives qui se prêtent à une mesure numérique d'un attribut donné. Mesurées selon une échelle de rapport, elles peuvent être de nature *continue* si elles sont susceptibles de prendre toutes les valeurs possibles à l'intérieur d'un intervalle donné (prix de vente, superficies), ou de nature *discrète* si elles ne prennent que des valeurs entières (âge de la propriété, nombre de salles de bain ou de chambres, nombre de places de stationnement, nombre de mois écoulés depuis la transaction);

- *les variables dichotomiques ou binaires (B)* utilisent une échelle nominale et servent essentiellement à désigner la présence (1) ou l'absence (0) d'un attribut. Très utilisées en analyse statistique, ces variables présentent l'avantage de ne requérir qu'une information

[1] En règle générale, la taille de la population se nomme *« N »* et celle de l'échantillon est *« n »*. Par souci de conformité avec le format du logiciel SPSS, *« N »* désignera tout au long du texte la taille de l'échantillon.

partielle sur les caractéristiques résidentielles. À titre d'exemple, une propriété possédant une piscine creusée sans que les dimensions de cette piscine ne soient connues; il sera possible, de même, de tenir compte d'un sous-sol fini sans en connaître la superficie. Pour mesurer la contribution de certains attributs, *une variable binaire multicatégorielle sera utilisée*, chaque catégorie étant comptabilisée comme une variable distincte par le modèle. C'est le cas, notamment, du *type d'immeubles* (« bungalow », cottage, unité jumelée ou en rangée) et *de la localisation* à l'intérieur du secteur d'analyse (sous-secteurs 1, 2, 3, etc.). Il est alors essentiel de déterminer *une catégorie de référence* (en général, la catégorie dominante) exclue du modèle, calibré par défaut pour cette catégorie. Ainsi, si le territoire d'étude comporte trois sous-secteurs et que le sous-secteur 1 est désigné comme référence, seuls les sous-secteurs 2 et 3 sont intégrés à l'analyse. Soulignons pour terminer que l'utilisation d'une variable binaire comme variable explicative exige que le nombre d'observations présentant cette caractéristique soit suffisant. D'une façon générale, *un minimum de 5 observations* sera requis pour intégrer une telle variable au modèle de régression. Dans certains cas toutefois, la force statistique du coefficient obtenu justifiera de maintenir la variable dans l'équation en dépit d'un nombre inférieur d'observations. La décision finale revient donc à l'analyste, qui devra cependant nuancer l'interprétation de ses résultats en conséquence.

- *les variables de rang (R)* sont des variables qualitatives utilisant une échelle ordinale et servant à ordonner les différentes valeurs prises par un attribut donné. Fort utiles pour décrire le niveau qualitatif d'une caractéristique résidentielle, à partir de l'évaluation subjective qu'en fait l'analyste (p. ex. : la qualité d'un aménagement paysager), il est approprié, cependant, de les utiliser avec circonspection. En effet, alors que la contribution marginale attribuée à la caractéristique, par le biais du coefficient de régression, sera proportionnelle à la valeur prise par la variable, soit le « nombre de points » accordé par l'analyste, il n'en va pas de même dans le marché, une cote « 2 » n'étant pas nécessairement équivalente au double d'une cote « 1 ». De plus, ce type de variable *ne doit pas servir à préjuger du comportement du marché*. Pour cette raison, il importe d'éviter de mesurer la contribution du parement extérieur, par exemple, par une variable de rang et lui substituer plutôt une variable binaire multicatégorielle. Enfin, il est nécessaire de rappeler que, pour les fins de l'analyse statistique, certaines données nécessitent d'être transformées. C'est le cas du code de classe où la qualité de la construction décroît en fonction de la croissance du code, alors que la cote accordée à la variable doit croître avec le degré de qualité de l'attribut (c.-à-d. la cote « 1 » désigne le niveau inférieur de qualité).

3.8.3.1 Exemple d'application

Le *tableau 3.1* de la page suivante présente la définition opérationnelle des 40 variables utilisées pour modéliser le marché des « bungalows ». Les notes explicatives accompagnant la définition des variables permettent de retracer la provenance de la variable, par rapport à la donnée apparaissant sur les formulaires 2.6.9 C ou 2.6.1 C, et la programmation utilisée pour adapter les données utiles à la méthode de comparaison.

Tableau 3.1 - Définition des variables utilisées

Nº	Nom	Définition opérationnelle	Type de variable
	Variable dépendante :		
	PRIXVTE	Prix de vente de l'immeuble vendu, en dollars	Métrique
	Variables indépendantes :		
1	VOISI_A	La propriété est située dans la famille d'unités de voisinage « A »	Binaire
2	VOISI_B	La propriété est située dans la famille d'unités de voisinage « B »	Binaire
3	VOISI_C	La propriété est située dans la famille d'unités de voisinage « C »	Binaire
4	AGEAPP	Âge apparent de la propriété, en nombre d'années	Métrique
5	LNAGEAPP	Logarithme naturel de l'âge apparent	Métrique
6	AIREHAB	Aire totale aux étages, en mètres carrés	Métrique
7	LNAIRHAB	Logarithme naturel de l'aire totale aux étages	Métrique
8	AIRFINSS	Aire finie au sous-sol, en mètres carrés	Métrique
9	CLASSUP	La propriété est de classe 4 ou moins	Binaire
10	PARDUR	Le parement extérieur est majoritairement composé de matériau dur	Binaire
11	FENRÉNOV	La fenestration de la propriété a fait l'objet d'une rénovation	Binaire
12	TOITRÉNO	La toiture de la propriété a fait l'objet d'une rénovation	Binaire
13	SDBAIN	Nombre de salles de bain complètes	Métrique
14	SDBSUP	Les salles de bain principales sont de qualité supérieure	Binaire
15	SDBEXTRA	Les salles de bain comportent au moins un extra	Binaire
16	SDEAU	Nombre de salles d'eau	Métrique
17	CUISSUP	La cuisine est d'une qualité supérieure	Binaire
18	CUISEXTR	La cuisine comporte au moins un extra	Binaire
19	CHAUFFÉL	La propriété est chauffée à l'électricité	Binaire
20	THERMOPO	La propriété possède une thermopompe	Binaire
21	ASPIR	La propriété possède un aspirateur central	Binaire
22	PROTECTI	La propriété est protégée par un système antivol	Binaire
23	PLAFCATH	La propriété possède un plafond cathédrale	Binaire
24	ENTRÉESS	La propriété a une entrée indépendante au sous-sol	Binaire
25	GRGALERI	La propriété possède une galerie supérieure à 15 m²	Binaire
26	NBFOYER	Nombre de foyers encastrés	Métrique
27	NBPLGAR	Nombre de places de garage	Métrique
28	ABRI	La propriété possède un abri d'auto	Binaire
29	NBPLSTAT	Nombre de places de stationnement	Métrique
30	PISCEXC	La propriété possède une piscine excavée	Binaire
31	PISCHT	La propriété possède une piscine hors terre	Binaire
32	SUPTER	Superficie de terrain en mètres carrés	Métrique
33	LNSUPTER	Logarithme naturel de la superficie du terrain	Métrique
34	TERRIRR	Le terrain a une forme irrégulière	Binaire
35	TERPENTE	Le terrain est en pente	Binaire
36	LOCACOIN	La propriété est localisée sur un site d'encoignure	Binaire
37	TTAMÉLOC	Le site bénéficie de toutes les améliorations locales possibles	Binaire
38	SECTPROG	Le secteur où est située la propriété est en progression	Binaire
39	MOIS	Nombre de mois écoulés entre la date de la vente et le premier juillet 1999	Métrique

Tableau 3.1 - Définition des variables utilisées (suite) –
Notes explicatives sur la programmation de certaines variables et la provenance des données

Nº	Nom	Champ du formulaire	Programmation
-	PRIXVTE	76PRIX[1]	
1	VOISI_A	00U[1]	
2	VOISI_B	00U[1]	
3	VOISI_C	00U[1]	
4	AGEAPP	00M[1]	= nombre d'années entre la date de transaction et la date de construction apparente
5	LNAGEAPP	AGEAPP[1]	Transformation logarithmique de AGEAPP
6	AIREHAB	01AFET[2]	
7	LNAIRHAB	AIREHAB[2]	Transformation logarithmique de AIREHAB
8	AIRFINSS	01AFSS[2]	
9	CLASSUP	00C[1]	Si 00C < 5, la variable prend la valeur 1; sinon, 0
10	PARDUR		Si (02PIERRE[2] + 02BRIQUE[2]) > 50, la variable prend la valeur 1; sinon, 0
11	FENRÉNOV	02OUVCDRENO[2]	Si 02OUVCDRENO = 1, la variable prend la valeur 1; sinon, 0
12	TOITRÉNO	02TOITCDRENO[2]	Si 02TOITCDRENO = 1, la variable prend la valeur 1; sinon, 0
13	SDBAIN	03_1SBNB[2]	
14	SDBSUP	03_1SBCDRENO[2]	Si 03_1SBCDRENO = 1, la variable prend la valeur 1; sinon, 0
15	SDBEXTRA	03_1SBEXTRANB[2]	Si 03_1SBEXTRANB > 0, la variable prend la valeur 1; sinon, 0
16	SDEAU	03_1SENB[2]	
17	CUISSUP	03_1CUISCDRENO[2]	Si 03_1CUISCDRENO = 1, la variable prend la valeur 1; sinon, 0
18	CUISEXTR	03_1CUISEXTRANB[2]	Si 03_1CUISEXTRANB > 0, la variable prend la valeur 1; sinon, 0
19	CHAUFFÉL	03_4CHAUFFCOMB[2]	Si 03_4CHAUFFCOMB = 3, la variable prend la valeur 1; sinon, 0
20	THERMOPO	03_5CLIMAT[2]	Si 03_5CLIMAT = 3, la variable prend la valeur 1; sinon, 0
21	ASPIR	03_6ASPICENT[2]	
22	PROTECTI	03_6PROTEC[2]	
23	PLAFCATH	03_6PLAFCATH[2]	
24	ENTRÉESS	03_6ENTREESS[2]	
25	GRGALERI	03_6GALERIE[2]	
26	NBFOYER	03_6NBFOYER[2]	
27	NBPLGAR		= (04G1NBPL[2] + 04G2NBPL[2])
28	ABRI	04AACDRENO[2]	Si 04AACDRENO > 0, la variable prend la valeur 1; sinon, 0
29	NBPLSTAT	04STNBPL[2]	
30	PISCEXC	04PISCICODE[2]	Si 04PISCICODE = 2, la variable prend la valeur 1; sinon, 0
31	PISCHT	04PISCICODE[2]	Si 04PISCICODE = 1, la variable prend la valeur 1; sinon, 0
32	SUPTER	79S[1]	
33	LNSUPTER	SUPTER[1]	Transformation logarithmique de SUPTER
34	TERRIRR	77D3[1]	
35	TERPENTE		Si 77E1 = 41, la variable prend la valeur 1; sinon, 0
36	LOCACOIN	77C5[1]	
37	TTAMÉLOC	77A8[1]	
38	SECTPROG	77V[1]	Si 77V = P, la variable prend la valeur 1; sinon, 0
39	MOIS	76DATE[1]	= (Nombre de mois écoulés entre la date de la vente (76DATE) et la date de référence (01-07-99) + 1)

[1] Données provenant de la fiche de propriété 2.6.9 C.
[2] Données provenant de la fiche de propriété 2.6.1 C.

3.9 ÉTAPE 4 : DESCRIPTION ET ANALYSE DE LA BASE DE DONNÉES

L'étape 4 aborde l'analyse statistique proprement dite des données. L'application des statistiques descriptives aux variables clés de la base de données constitue une étape préalable indispensable, et trop souvent négligée, à la modélisation. Elle permet, en effet, de mieux comprendre la nature du marché immobilier à l'étude et d'en visualiser la structure. C'est notamment à cette étape que l'analyste pourra juger de la pertinence de procéder à des transformations mathématiques de certaines variables. L'analyse des distributions statistiques permet enfin de déceler les cas problèmes (valeurs extrêmes ou aberrantes) et, le cas échéant, les erreurs ou omissions majeures dans la saisie de l'information.

Les indicateurs statistiques utilisés comprennent *les indicateurs de tendance centrale* (moyenne, médiane et mode), *les indicateurs de dispersion* (écart type, étendue, minimum et maximum) et *les indicateurs de forme* (coefficients d'asymétrie « skewness » et d'aplatissement « kurtosis »); plus rarement utilisés et facultatifs, ces derniers complètent, néanmoins, le portrait statistique des attributs sélectionnés. Il est, par ailleurs, essentiel d'accompagner le tableau des résultats d'une représentation graphique.

La base utilisée ici est structurée en format SPSS; il est également possible de la travailler sur fichier Excel et de l'importer ensuite dans SPSS. Une fois ouvert le fichier SPSS contenant les données, on distingue la barre des menus dont le quatrième, intitulée « ***Data*** », permet de structurer et de modifier la base. À l'intérieur de ce menu apparaissent plusieurs commandes. En particulier… :

- la commande « ***Define Variable*** » permet de nommer la variable, d'en définir le format ainsi que l'échelle de mesure. On peut également lui attribuer un label qui en facilite l'identification et choisir une procédure de substitution pour les données manquantes;

- la commande « ***Insert Variable*** » permet d'insérer une colonne additionnelle à la gauche de la variable sélectionnée dans la base;

- la commande « ***Sort cases*** » permet de ranger une variable donnée par ordre croissant ou décroissant;

- la commande « ***Split file*** » permet de scinder l'analyse en fonction de la variable de subdivision sélectionnée. À titre d'exemple, si le territoire d'étude comporte trois secteurs spatiaux identifiés dans la base de données par un code numérique (i.e. 1, 2 et 3), le choix du secteur comme variable de subdivision aura pour effet de générer autant d'analyses qu'il y a de secteurs;

- la commande « ***Select cases*** » permet notamment de créer une variable filtre de valeur 0 ou 1 (cocher l'item « ***Use filter variable*** ») qui servira à pointer les observations que l'on veut retirer de l'analyse (0). Cette procédure est notamment utile pour identifier les transactions devant faire l'objet du test de validation final (voir la sous-section 2.3.2) et pour retirer les résidus extrêmes de l'analyse (Étape 7).

Le menu suivant, dénommé « ***Transform*** », comporte également plusieurs commandes, dont… :

- la commande « ***Compute variable*** », utilisée pour créer une variable cible (« ***Target Variable*** ») préalablement définie (colonne vide) par transformation mathématique d'une autre variable figurant déjà dans la base. La nouvelle variable (e.g. LnSUPHAB) est alors

crée en plaçant dans la zone «***Numeric Expression*** » la variable de départ (SUPHAB), transformée par la fonction («***Functions*** ») appropriée (LN). La même procédure est utilisée pour générer une variable interactive en effectuant le produit ou le ratio de deux variables existantes;

- la commande « ***Replace missing values*** » offre plusieurs options de remplacement des données manquantes, ce qui permet de conserver les observations pour lesquelles l'information est incomplète sans pour autant biaiser les résultats des analyses.

3.9.1 APPLICATION DES STATISTIQUES DESCRIPTIVES

L'application des statistiques descriptives permet l'étude des caractéristiques numériques de regroupements d'observations. Elle fournit les moyens de décrire ces ensembles et de résumer l'information recueillie sur les éléments les constituant.

3.9.1.1 Exemple d'application

L'analyse se limite ici aux variables quantitatives (métriques) se prêtant à une représentation des distributions statistiques. Six caractéristiques sont retenues, soit le prix de vente (PRIXVTE), le prix unitaire au mètre carré (PRIXM2), l'âge apparent (AGEAPP), l'aire habitable (AIREHAB), la superficie du terrain (SUPTER) et le nombre de mois écoulés entre la date de référence (01-07-99) et la date de transaction (MOIS). L'analyse porte, dans un premier temps, sur l'ensemble des transactions *bona fide*, soit 147 propriétés. Les résultats de l'analyse apparaissent au *tableau 3.2* de la page suivante.

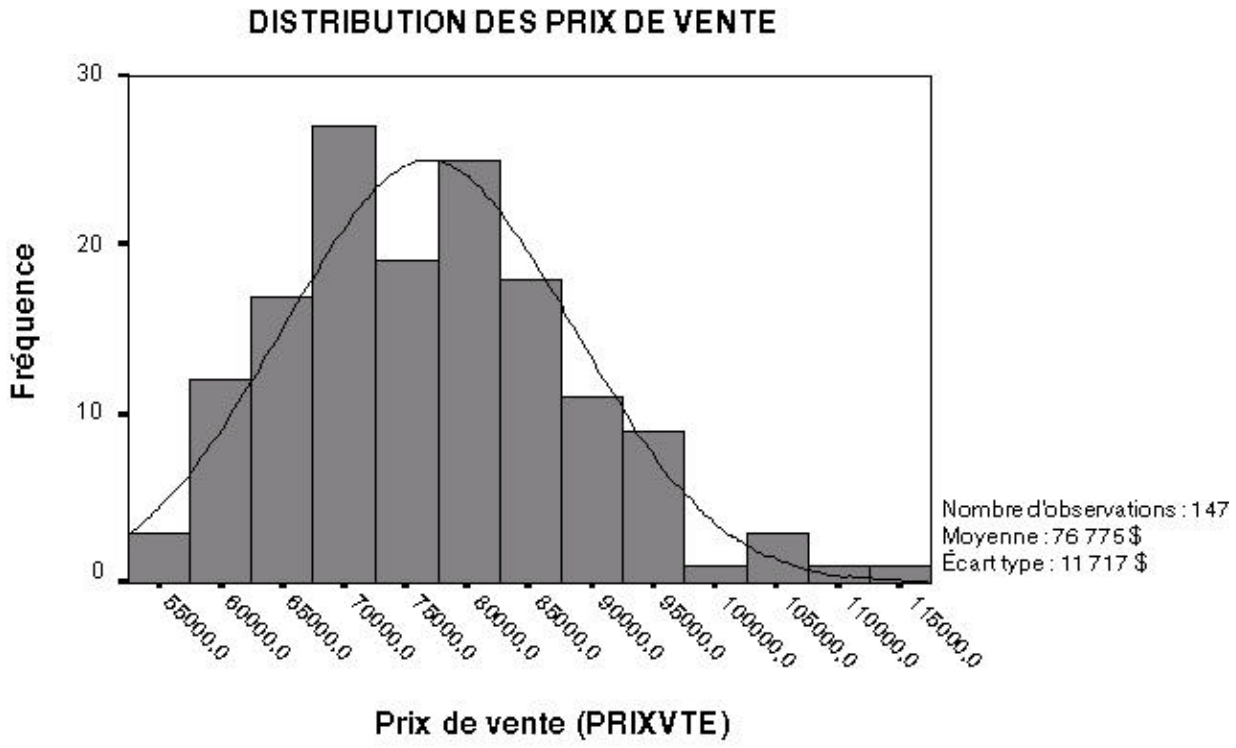
Comme il est possible de le constater, la variable dépendante (PRIXVTE) présente une distribution se rapprochant de la normale, les prix moyen (76 775 $) et médian (76 000 $) étant très similaires. Tel qu'indiqué dans la note sous le tableau, une distribution bimodale est représentée, dont seul l'indicateur de valeur inférieure (65 000 $) apparaît ici. L'écart type des prix (11 717 $ ou 15,3 %[1]) est relativement faible alors que l'étendue de la distribution (62 000 $) est très raisonnable. Enfin, la distribution est quelque peu décalée vers la droite (coefficient d'asymétrie de 0,57 versus 0 pour la normale) et présente une forme à peine plus aiguë que la normale (coefficient d'aplatissement de 0,46 versus 0 pour la normale). Bref, la distribution des prix de vente *peut être assimilée à une distribution normale*, ce qui est de bonne augure pour la suite des choses. En effet, si la condition de normalité n'est pas essentielle, en ce qui a trait aux variables explicatives du modèle (par définition, la distribution des variables binaires n'est pas normale), elle est par contre incontournable dans le cas de la variable dépendante.

[1] 15,3 % représente le coefficient de variation résultant du rapport entre 11 717 $ et 76 775 $.

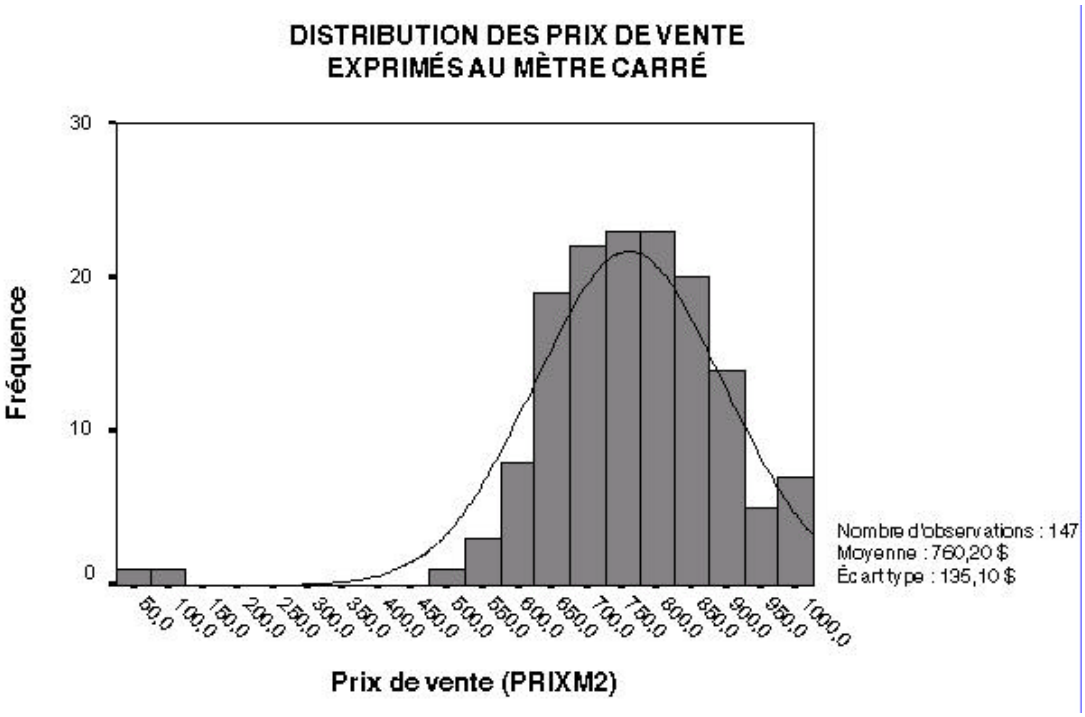
Tableau 3.2 - Statistiques descriptives -
Ensemble de l'échantillon avant corrections (N=147)

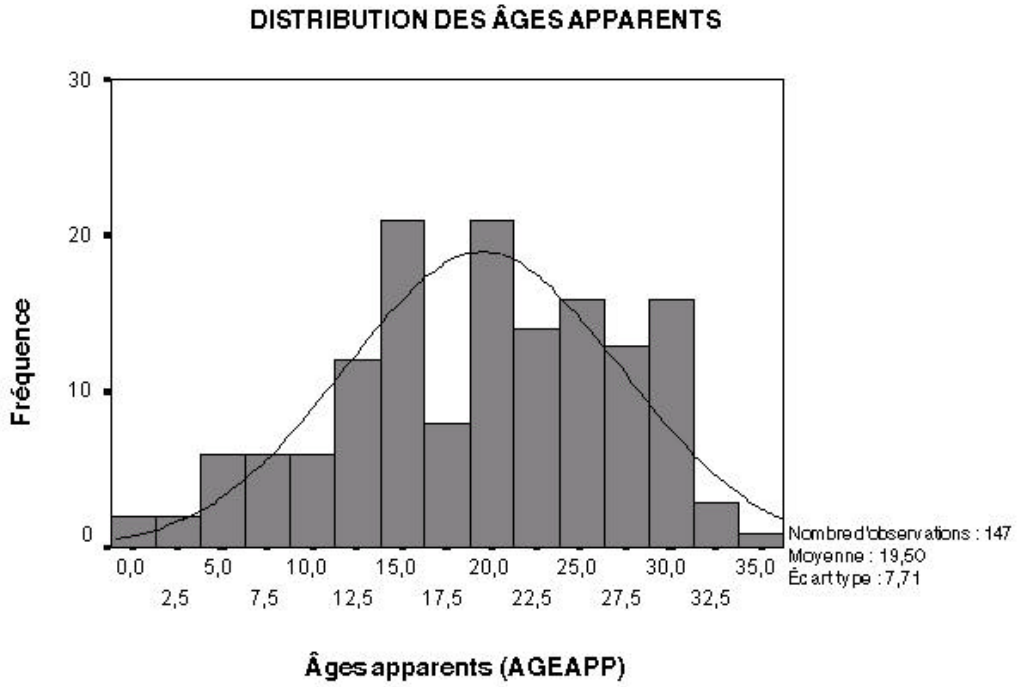
	PRIXVTE	PRIXM2	AGEAPP	AIREHAB	MOIS	SUPTER
Nombre d'observations	147	147	147	147	147	147
Moyenne	$76,775	$760	19	112	21	865
Médiane	$76,000	$763	20	99	19	750
Mode	$65,000	$620	16	97	30	836
Écart type	$11,717	$135	8	99	6	434
Coefficient d'asymétrie	,57	-1,57	-,37	8,30	,08	4,25
Écart d'asymétrie	,20	,20	,20	,20	,20	,20
Coefficient d'aplatissement	,46	7,30	-,60	69,01	-,10	28,06
Écart d'aplatissement	,40	,40	,40	,40	,40	,40
Étendue	$62,000	$944	33	897	29	3884
Valeur minimale	$54,000	$74	1	73	3	446
Valeur maximale	$116,000	$1,018	34	970	32	4330

1 Il existe plusieurs modes, mais seule la valeur inférieure apparaît dans ce tableau.

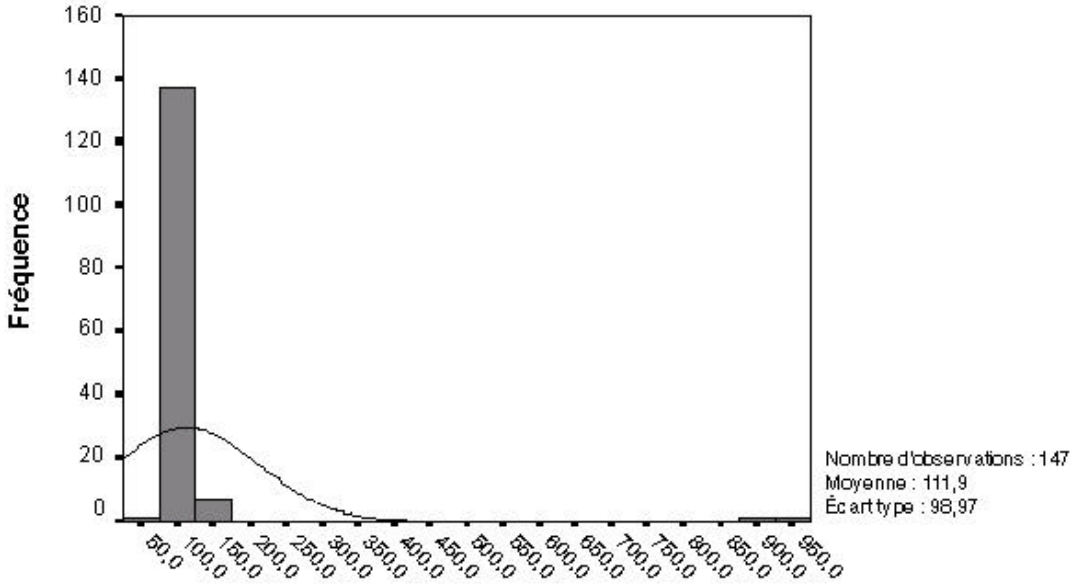


L'analyse de la seconde variable (PRIXM2) suggère déjà une irrégularité dans les données. Le graphique de la page suivante montre en effet une distribution non seulement très pointue (coefficient d'aplatissement de 7,30), mais surtout fortement décalée vers la gauche (coefficient d'asymétrie de -1,57) en raison d'un minimum très faible ($74/m²) qui semble aberrant. L'aberration se confirme à l'analyse de la variable AIREHAB dont la distribution laisse apparaître deux valeurs extrêmes, de plus de 900 mètres carrés, ayant pour effet de gonfler indûment l'écart type de la variable. Il s'agit là d'un exemple éloquent d'erreurs de saisie qui, si elles ne sont pas corrigées, viendront fausser de façon substantielle les résultats de l'analyse de régression.

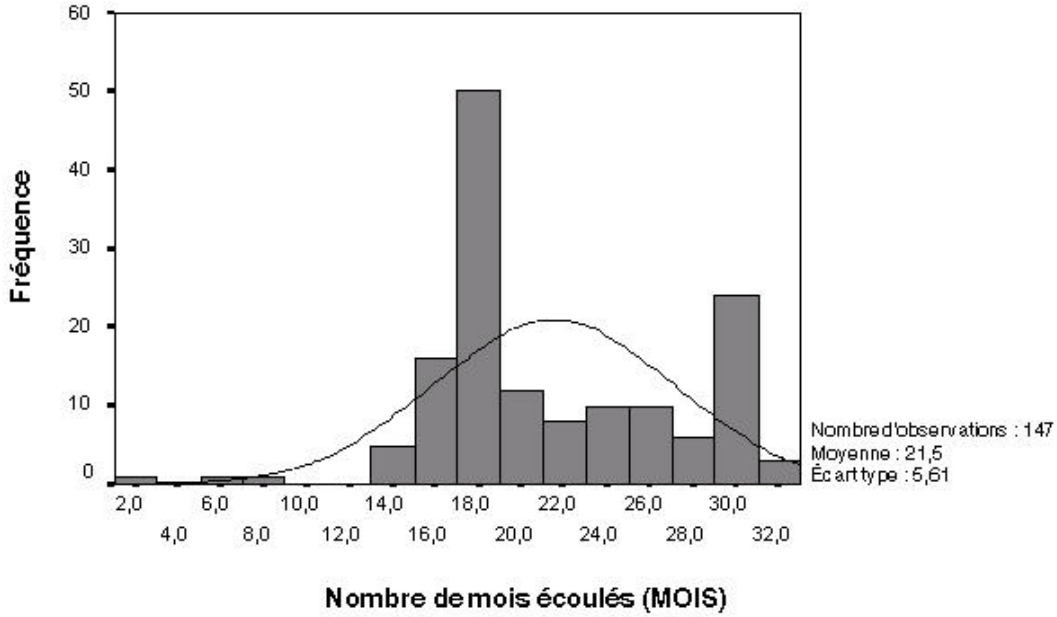

DISTRIBUTION DES PRIX DE VENTE
EXPRIMÉS AU MÈTRE CARRÉ
Fréquence
30
20
10
0
50,0
100,0
150,0
200,0
250,0
300,0
350,0
400,0
450,0
500,0
550,0
600,0
650,0
700,0
750,0
800,0
850,0
900,0
950,0
1000,0
Prix de vente (PRIXM2)
Nombre d'observations : 147
Moyenne : 760,20 $
Écart type : 195,10 $


DISTRIBUTION DES ÂGES APPARENTS
Fréquence
30
20
10
0
0,0
2,5
5,0
7,5
10,0
12,5
15,0
17,5
20,0
22,5
25,0
27,5
30,0
32,5
35,0
Âges apparents (AGEAPP)
Nombre d'observations : 147
Moyenne : 19,50
Écart type : 7,71

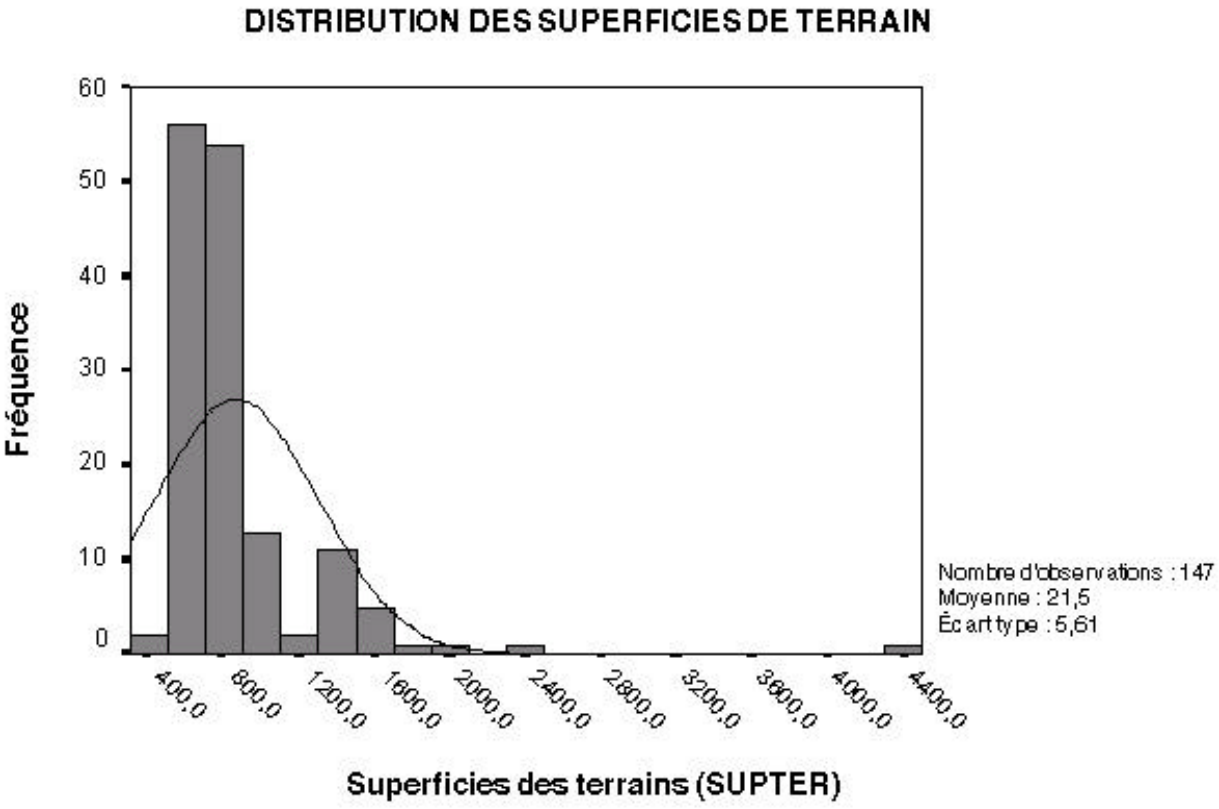
DISTRIBUTION DES AIRES HABITABLES



DISTRIBUTION DU TEMPS ÉCOULÉ ENTRE LA DATE DE TRANSACTION ET LA DATE DE RÉFÉRENCE



Le graphique suivant montre la variable « SUPTER » (superficie du terrain), force est de constater qu'en plus d'être particulièrement concentrée, sa distribution présente un décalage marqué vers la droite, en raison d'une valeur extrême (4 330 m^2), ce qui a pour effet de hausser la moyenne et son écart type. Cette réalité du marché permet d'envisager de transformer mathématiquement la variable pour tenir compte de l'étendue des valeurs.



3.9.2 TRANSFORMATION MATHÉMATIQUE DE VARIABLES

Dans la mesure où le recours à une méthode linéaire d'analyse de régression (méthode des moindres carrés ordinaires) est possible, le traitement de relations non linéaires, phénomène fréquent en évaluation immobilière, implique, au préalable, que la variable visée soit transformée. C'est notamment le cas des aires des constructions, des superficies de terrain (surtout) et de l'âge de la propriété dont l'étendue des valeurs peut, dans certains marchés, être très prononcée. Dans de telles situations, la contribution unitaire de la variable (exprimée en $/m² ou en $/an) ne peut plus être assimilée à une constante. Plusieurs transformations peuvent être utilisées. En évaluation de masse, les plus fréquemment utilisées sont[1] :

- ***la transformation réciproque***, consistant à remplacer les valeurs attribuées à la variable par leur inverse (1/X); fondée sur le principe de contribution marginale décroissante, cette transformation se prête bien à la prise en compte, par exemple, de l'effet de la distance au centre-ville sur les valeurs marchandes;

[1] IAAO, op. cit., p. 321-325.

- ***la transformation exponentielle***, consistant à élever les valeurs d'une variable à une certaine puissance de façon à accroître l'écart les séparant (exposant >1) ou à le comprimer (exposant compris entre « 0 » et « 1 »). Le recours à un exposant négatif ramène à la transformation réciproque. Les transformations exponentielles les plus fréquemment utilisées demeurent l'élévation au carré (X^2) et la racine carré (X, ou $X^{0,5}$). Cette dernière est souvent appliquée à la superficie du terrain, lorsque l'échantillon comporte des terrains de dimensions très variables dont la recherche consiste à réduire les écarts. Comme la transformation réciproque, la transformation exponentielle repose sur le principe de contribution marginale décroissante;

- ***la transformation logarithmique***, l'une des plus utilisées, permet de linéariser la relation propre à une distribution statistique très étendue et peut être exprimée sur la base de logarithmes décimaux (Log) ou de logarithmes naturels (Ln). Appliquée simultanément à la variable dépendante, soit le prix de vente, et aux variables indépendantes de type métrique, la transformation logarithmique permet de générer un *modèle multiplicatif*, dont les termes de l'équation sont multiplicatifs plutôt qu'additifs. Outre le fait qu'un tel modèle tend à normaliser la distribution des prix de vente, les coefficients de régression qu'il génère agissent comme des facteurs d'ajustement supérieurs ou inférieurs à l'unité plutôt que comme des prix unitaires fixes[1];

- ***la transformation multiplicative*** repose sur le principe d'interaction entre les attributs et consiste à créer une variable interactive à partir du produit de deux variables distinctes, dont l'une est une variable binaire ou de rang (p. ex. : AIREHAB * CLASSUP). Une telle transformation peut, notamment, être utilisée pour apporter plus de nuance à l'établissement de la contribution marginale d'un attribut majeur, comme l'aire habitable, lorsque le même modèle regroupe plusieurs types de propriétés (p. ex. : « bungalow » et cottage) ou plusieurs qualités de construction (p. ex. : « condominiums » de basse, moyenne et haute densité). Un prix unitaire ($/m²) par catégorie est ainsi obtenu;

- ***la transformation « ratio »*** enfin, est une variante de la précédente et consiste à diviser une variable par une autre. Le prix au m² (PRIXVTE / AIREHAB), souvent utilisé pour contrer les effets de l'hétéroscédasticité (voir le point 3.4.4.6 de ce chapitre), est un bon exemple.

Ces transformations peuvent être utilisées seules ou combinées les unes aux autres. Dans la suite de l'analyse, une transformation logarithmique est appliquée à certaines variables du modèle statistique, y compris à la variable dépendante. Le recours aux transformations sur les variables est donc une façon élégante de tenir compte des relations non linéaires entre le prix et les caractéristiques des immeubles; il permet d'améliorer la performance et la finesse prédictive du modèle sans avoir à recourir aux méthodes de régression non linéaires, plus complexes à manipuler et plus difficiles d'interprétation.

En conclusion à cette première partie de l'analyse descriptive, la distribution des variables « AGEAPP » et « MOIS » ne soulève aucun problème particulier. L'âge apparent présente une distribution bimodale assez prononcée (à 16 et 21 ans), alors que l'indice d'inflation (MOIS) montre très clairement que la majorité des transactions se concentre autour des mois de juin 1998 et 1999.

[1] Voir à cet effet : IAAO, op. cit., p. 388-389.

3.9.3 APPLICATION DES STATISTIQUES DESCRIPTIVES APRÈS CORRECTIONS ET ÉPURATION

Les mêmes statistiques descriptives sont appliquées à l'échantillon, mais cette fois-ci après y avoir retiré le sous-échantillon identifié pour la validation ultérieure et les corrections apportées à la base de données, à la suite de la détection d'erreurs de saisie. De plus, la transaction n^{O} 82 est retirée de l'échantillon puisqu'elle est la seule qualifiée comme étant de classe inférieure. La règle d'usage exige au moins quatre transactions de ce type pour les conserver dans l'échantillon servant à déterminer les variables indépendantes.

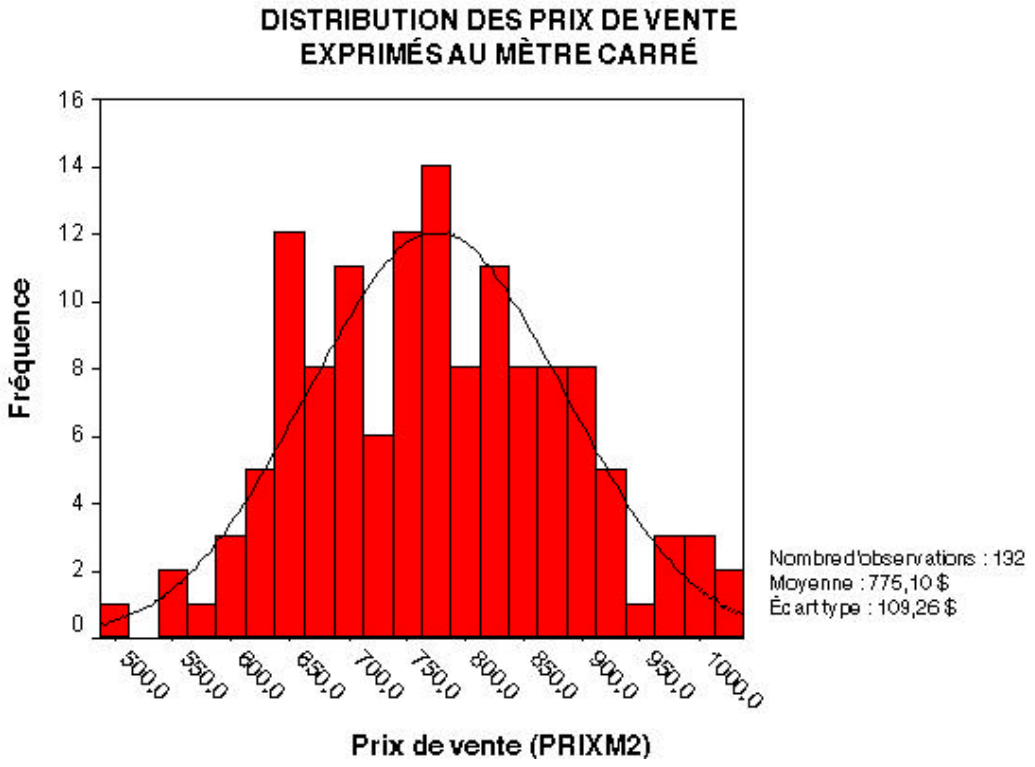
3.9.3.1 Exemple d'application

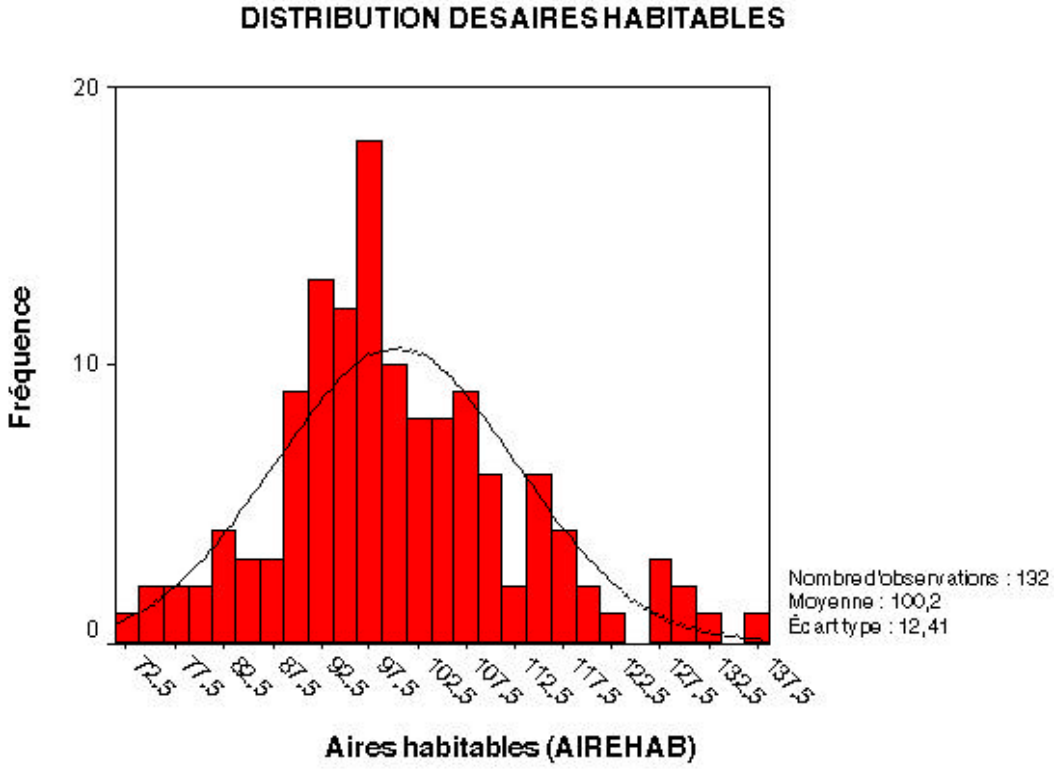
Le *tableau 3.3* reproduit la nouvelle structure de la base de données, après corrections des erreurs identifiées plus haut et après extraction du sous-échantillon destiné à la validation du modèle final. L'analyse porte sur 132 observations et démontre, dans le cas de l'aire habitable notamment, une nette réduction de sa dispersion autour de la moyenne. Ainsi, le coefficient de variation[1] de cette variable clé est passé de 88,4 % avant correction à 12 % après correction. Les coefficients d'asymétrie et d'aplatissement se sont normalisés alors que l'étendue de la distribution s'est considérablement réduite, passant de 897 m² à 65 m².

Tableau 3.3 - Statistiques descriptives - Ensemble de l'échantillon après corrections (N=132)

	PRIXVTE	PRIXM2	AGEAPP	AIREHAB	MOIS	SUPTER
Nombre d'observations	132	132	132	132	132	132
Moyenne	$77,253	$775	$20	$100	$21	$866
Médiane	$76,750	$770	$20	$98	$19	$744
Mode	$72,000	$620	$16	$90	$30	$502
Écart type	$11,969	$109	$8	$12	$6	$450
Coefficient d'asymétrie	,50	,10	-,31	,55	,07	4,24
Écart d'asymétrie	,21	,21	,21	,21	,21	,21
Coefficient d'aplatissement	,39	-,49	-,60	,51	-,07	27,01
Écart d'aplatissement	,42	,42	,42	,42	,42	,42
Étendue	$62,000	$510	33	65	29	3884
Valeur minimale	$54,000	$508	1	73	3	446
Valeur maximale	$116,000	$1,018	34	138	32	4330

[1] Le coefficient de variation (CV) d'une variable se définit comme le ratio de son écart type par sa moyenne.


DISTRIBUTION DES PRIX DE VENTE
EXPRIMÉS AU MÈTRE CARRÉ
Fréquence
16
14
12
10
8
6
4
2
0
500,0
550,0
600,0
650,0
700,0
750,0
800,0
850,0
900,0
950,0
1000,0
Prix de vente (PRIXM2)
Nombre d'observations : 132
Moyenne : 775,10 $
Écart type : 109,26 $


DISTRIBUTION DES AIRES HABITABLES
Fréquence
20
10
0
72,5
77,5
82,5
87,5
92,5
97,5
102,5
107,5
112,5
117,5
122,5
127,5
132,5
137,5
Aires habitables (AIREHAB)
Nombre d'observations : 132
Moyenne : 100,2
Écart type : 12,41

3.10 ÉTAPE 5 : ANALYSE DE CORRÉLATION

Il arrive fréquemment, en analyse statistique, d'émettre l'hypothèse qu'il puisse exister entre deux variables un lien quelconque. C'est notamment le cas en évaluation immobilière où la recherche consiste à établir une relation linéaire entre, d'une part, la valeur d'une unité d'évaluation et, d'autre part, son aire habitable, son âge, sa localisation, les revenus qu'elle génère ou toute autre caractéristique pertinente. Ici, la notion de « *corrélation* » est abordée. Il y a corrélation entre deux variables observées, sur les éléments d'une même population, lorsque les valeurs prises par les deux variables fluctuent simultanément, soit dans le même sens (corrélation positive), soit en sens contraire (corrélation négative).

À cette étape, l'évaluateur cherche, par un test de corrélation simple, à identifier les variables ne présentant aucune ou peu de corrélation avec le prix de vente, afin de les soustraire de l'analyse de régression.

3.10.1 TEST DE CORRÉLATION SIMPLE (*r*)

Dans le cas de la *corrélation simple*, à laquelle se limite ce texte, l'analyse du phénomène consiste à prélever, sur une population, un échantillon aléatoire de taille « *n* », et d'observer, sur chaque unité de cet échantillon, les valeurs de deux variables statistiques nommées « X » et « Y ». Cet échantillon contient ainsi « *n* » couples (x_i, y_i). Le coefficient de corrélation linéaire simple[1] , noté « *r* », est un nombre variant entre « +1 » et « -1 » et mesurant l'intensité de la relation linéaire entre deux variables. Un coefficient se rapprochant de « +1 » indique une forte relation linéaire *positive* entre « X » et « Y » alors que cette relation est tout aussi forte, mais *négative*, si le coefficient tend vers « -1 ». Quant à l'absence de relation linéaire entre « X » et « Y », elle se traduit par un coefficient de corrélation qui tend vers zéro. Il est important de souligner ici que l'absence de relation *linéaire* entre deux variables ne signifie pas, malgré cela, qu'une relation de nature *non linéaire* ne puisse exister entre elles. Si tel est le cas, cette relation ne sera connue qu'en appliquant, au préalable, à l'une ou, le cas échéant, aux deux variables, la transformation mathématique pertinente. Enfin, si l'amplitude et le signe du coefficient de corrélation simple sont importants pour qualifier la force d'une relation entre deux variables, son *degré de signification statistique* l'est tout autant.

3.10.2 TEST DE FIABILITÉ (H_0)

Comme il est d'usage en inférence statistique, tout résultat doit être accompagné d'un test de fiabilité indiquant la probabilité qu'une telle occurrence ne relève du simple hasard2, ce qu'il est convenu d'appeler *l'hypothèse nulle*, ou H_O. Dans le cas concerné, il s'agit d'un *test bilatéral* (two-tailed test) dont la valeur, c.-à-d. la probabilité d'occurrence, est d'autant plus faible que la relation entre les deux variables est fiable. Les *seuils de signification statistique* les plus couramment utilisés en statistique sont ceux de 0,01 (1 %), de 0,05 (5 %) et de 0,10 (10 %). Le seuil de *0,05* est de loin le plus utilisé et constitue la norme, retenue par défaut sur la plupart des logiciels statistiques. Il indique que dans 5 % des cas, ou une fois sur vingt, la corrélation

[1] En l'occurrence, il s'agit ici du coefficient de Pearson, le coefficient de Spearman étant utilisé pour traiter des variables de rang.

[2] La théorie des probabilités rappelle, en effet, que si tout est possible, rien n'est certain. Cela se reflète bien non seulement dans la définition de la valeur marchande, mais aussi dans l'existence humaine, où les deux seules certitudes sont la mort et les impôts.

obtenue est fortuite et que l'analyste commet donc une erreur en l'interprétant comme une indication de la force réelle de la relation entre « X » et « Y[1] » . Si un seuil de 0,05 est considéré comme acceptable en recherche, lequel correspond à un *intervalle de confiance* de 95 %, une probabilité supérieure à ce seuil devient un motif suffisant de rejet du résultat obtenu (le coefficient de corrélation) et, par conséquent, d'acceptation de l'hypothèse nulle (H_0), en vertu de laquelle le coefficient n'est pas statistiquement significatif. En d'autres termes, il n'est pas significativement différent de zéro et ne peut donc être interprété sans un risque excessif d'erreur.

Dans la procédure de modélisation, l'analyse de corrélation constitue une étape intermédiaire entre la description et l'analyse de la base de données (étape 4) et l'analyse de régression proprement dite (étape 6). Comme il est mentionné plus haut, la matrice des corrélations simples ne permet pas d'établir de liens de cause à effet entre les diverses variables. Par ailleurs, elle ne prend pas en considération les influences croisées entre ces variables, comme le fait l'analyse de régression, et peut même donner lieu à de fausses interprétations lorsqu'il y a présence de forte multicolinéarité. Elle demeure, néanmoins, un outil fort utile pour se faire une première idée des attributs résidentiels ayant, potentiellement, le plus d'impact sur les prix, tout en contribuant à détecter la multicolinéarité excessive. Une analyse judicieuse des corrélations renseigne également sur la nature du parc et, éventuellement, sur le profil des résidants du secteur.

3.10.3 EXEMPLE D'APPLICATION

Dans l'exemple de la page suivante (*tableau 3.4*), seulement cinq variables ont été retenues sur les quelque 40 disponibles dans la banque de données utilisée et ce, de façon à ne pas alourdir inutilement le tableau. Dans la mesure où il s'agit de corrélations simples, donc entre des variables prises deux à deux, les influences croisées ne sont pas prises en compte ici et les coefficients seraient identiques si la matrice des corrélations englobait l'ensemble des variables de la base de données. Tout d'abord, la matrice en question, une matrice carrée de dimension 5 x 5, est symétrique par rapport à la diagonale, les éléments du triangle supérieur droit étant reproduits exactement dans la portion inférieure gauche du tableau. Par ailleurs, chaque cellule comporte trois termes : le premier est le coefficient de corrélation, le second représente la probabilité de l'erreur et le troisième indique la taille de l'échantillon « *N* » utilisé dans chacun des cas (ici, 132 observations).

[1] Ceci repose sur l'hypothèse où l'analyse est refaite un très grand nombre de fois, les résultats différant évidemment à chaque tentative.

Tableau 3.4 - Matrices des corrélations (5 variables/132 observations)

		PRIXVTE	AGEAPP	AIREHAB	AIRFINSS	SUPTER
PRIXVTE	Coefficient de corrélation	1,000	-,403**	,513**	,362**	,295*
	Degré de signification	,	,000	,000	,000	,001
	Nombre d'observations	132	132	132	132	132
AGEAPP	Coefficient de corrélation	-,403**	1,000	,172*	-,125	-,121
	Degré de signification	,000	,	,049	,152	,167
	Nombre d'observations	132	132	132	132	132
AIREHAB	Coefficient de corrélation	,513**	,172*	1,000	,116	,143
	Degré de signification	,000	,049	,	,185	,101
	Nombre d'observations	132	132	132	132	132
AIRFINSS	Coefficient de corrélation	,362**	-,125	,116	1,000	,040
	Degré de signification	,000	,152	,185	,	,652
	Nombre d'observations	132	132	132	132	132
SUPTER	Coefficient de corrélation	,295**	-,121	,143	,040	1,000
	Degré de signification	,001	,167	,101	,652	,
	Nombre d'observations	132	132	132	132	132

**. La corrélation est significative à partir du seuil de 0,01.

*. La corrélation est significative à partir du seuil de 0,05.

Les coefficients de corrélation marqués d'un astérisque sont statistiquement significatifs au seuil de 0,05, alors que ceux marqués d'un double astérisque le sont au seuil de 0,01. À la première colonne du tableau, présentant les corrélations entre le prix de vente, d'une part, et les quatre autres variables sélectionnées, d'autre part, les coefficients sont tous relativement élevés et fortement significatifs (probabilité d'erreur de 0,001 ou moins). Il est possible de croire que la plus forte relation linéaire concerne l'aire habitable (0,513), alors que le prix et l'âge apparent de l'immeuble sont négativement reliés (-0,403). Il existe également une forte relation positive entre le prix de vente, l'aire finie du sous-sol (0,362) et la superficie du terrain (0,295). Des autres corrélations de la matrice, seule celle relative au couple « AIREHAB-AGEAPP » (0,172) est statistiquement significative au seuil de 5 % (0,049).

En conclusion, les cinq variables analysées sont statistiquement corrélées avec le prix de vente et doivent être conservées pour l'analyse de régression à la prochaine étape.

3.11 ÉTAPE 6 : ANALYSE DE RÉGRESSION

Pour cette étape, deux procédures de régression seront utilisées, soit la procédure de régression par étape automatisée (procédure *Stepwise Regression*) et l'approche standard (procédure *Enter*) comportant les quatre sous-étapes suivantes :

1. Produire une première équation en utilisant l'ensemble des variables retenues à l'étape 3 (voir le point 3.8 de ce chapitre);

2. Réduire substantiellement le nombre de variables explicatives en éliminant celles ne répondant pas aux critères statistiques retenus de façon à améliorer la performance globale du modèle;

3. Chercher à améliorer la performance de l'équation en transformant mathématiquement certaines variables indépendantes;

4. Vérifier les performances du modèle multiplicatif par rapport au modèle additif en appliquant une transformation logarithmique au prix de vente.

3.11.1 PROCÉDURE DE RÉGRESSION STANDARD

L'application de la procédure de régression standard consiste à « forcer » dans l'équation toutes les variables souhaitées. En pratique, il revient évidemment à l'analyste de choisir lui-même les caractéristiques qui lui semblent les plus pertinentes sur la base de son expérience et de sa connaissance du marché local. Il importe, toutefois, d'éviter de trop restreindre la sélection, par crainte d'éliminer *a priori* des attributs résidentiels dont seule l'analyse peut dire s'ils contribuent ou non à la formation des valeurs marchandes. La prudence commande plutôt d'amorcer l'analyse avec un nombre relativement élevé de variables, puis d'en réduire progressivement le nombre selon une procédure itérative, sur la base des tentatives précédentes.

Pour actionner cette procédure, il suffit d'entrer dans le menu « ***Analyze*** », puis d'actionner la commande « ***Regression*** » et la sous-commande « ***Linear*** ». On choisit alors la variable dépendante (PRIXVTE) et les variables indépendantes à même la liste de gauche. Dans la mesure où l'on a préalablement défini dans le menu « ***Data*** » une variable filtre qui sert à l'ensemble des traitements statistiques[1], le recours à l'item « ***Selection variable*** » devient superflu. L'utilisateur active ensuite l'onglet « ***Statistics*** » qui offre plusieurs possibilités. Pour les besoins de la démonstration, nous avons coché les items « ***Estimates*** », « ***Model fit*** » et « ***Collinearity diagnostics*** » qui fournissent respectivement les paramètres de la régression et leurs tests, les indicateurs de performance du modèle ainsi que les valeurs VIF (Variance Inflation Factors) qui permettent de détecter la multicollinéarité excessive. Bien que fort utiles pour l'interprétation des coefficients de régression individuels, les intervalles de confiance des paramètres B_i (item « ***Confidence intervals*** ») n'ont pas été retenus afin de ne pas surcharger la présentation des résultats. Quant à l'analyse des résidus (« ***Residuals*** ») et aux graphiques connexes (onglet « ***Plots*** »), nous en gardons l'interprétation pour plus tard.

1 Dans le cas qui nous concerne, c'est la variable « *delete* » qui est utilisée à cette fin.

3.11.1.1 Indicateurs de performance et les tests d'hypothèse

Les principaux indicateurs de performance globale du modèle se définissent comme suit :

- ***le coefficient de corrélation multiple (R)*** : Ce coefficient reflète la force de la relation linéaire entre la variable dépendante « Y » (PRIXVTE) et l'ensemble des variables indépendantes du modèle;

- ***le coefficient de détermination « R² », ou « R carré »* (*R Square*)** : Le coefficient de détermination est un indicateur de la performance *explicative* du modèle. Il représente la proportion de la variation totale de « Y » expliquée par l'ensemble des variables explicatives de l'échantillon;

- ***le « R² » ajusté*[1]** : Bien que l'ajout de variables explicatives dans l'équation de régression ait pour effet de faire augmenter le « R² », il n'entraîne pas nécessairement une augmentation du pouvoir explicatif du modèle. En effet, lorsque le nombre de variables explicatives (*k*) devient relativement grand par rapport à la taille de l'échantillon (*N*)[1], la contribution marginale d'un attribut additionnel peut ne pas compenser la perte d'un *degré de liberté*, le nombre de degrés de liberté étant fonction de l'écart entre *« N »* et *« k »*. Alors que le nombre de degrés de liberté n'est pas pris en compte par le « R² », il l'est par le « R² » ajusté : au fur et à mesure que *« k »* augmente, ce dernier atteint un maximum, correspondant au nombre optimal de variables explicatives à inclure dans le modèle, puis décroît par la suite;

- ***l'erreur type d'estimation (Standard Error of Estimate, ou SEE)*** est l'indicateur de la performance *prédictive* du modèle de régression. Exprimé en dollars dans la forme linéaire du modèle, cet indicateur représente l'erreur commise dans l'estimation des prix de vente par le modèle. Il est cependant plus pratique d'exprimer cette erreur en pourcentage du prix moyen de l'échantillon; *l'erreur type relative est alors obtenue*, ou *coefficient de variation* (*CV*). Selon les normes de l'IAAO, un modèle est performant si le CV est de 15 % ou moins. Dans un modèle d'évaluation où la performance prédictive est cruciale, le seuil de 10 % sera plutôt visé, ce qui correspond à l'erreur type relative moyenne générée par les modèles statistiques;

- **le Test de Fisher (*F*)** constitue un test d'hypothèse afin de vérifier si les variables indépendantes contribuent à expliquer la variable dépendante. Pour un seuil de signification statistique de 0,05, la valeur critique de « F » pour 38 degrés de liberté au numérateur (df_1 = k = 38) et 93 degrés de liberté au dénominateur (df_2 = $N - k - 1$ = 93) est d'environ 1,6. La valeur du « F » calculé pour l'échantillon doit donc dépasser cette valeur critique pour que le modèle soit significatif à ce niveau;

- **l'erreur type du coefficient B_i** non standardisé, c.-à-d. originel, de la régression *(Standard Error)*. Pour illustrer le rôle de ce dernier indicateur, il est permis de constater que l'erreur type est au coefficient de régression ce que l'écart type est à la moyenne d'une variable : *une mesure de sa variabilité*. Ainsi, si plusieurs régressions étaient effectuées sur autant de sous-échantillons tirés d'un même échantillon principal, la valeur des paramètres de régression ainsi obtenus différerait d'une fois à l'autre. La « robustesse » d'un coefficient

[1] Voir à cet effet : Martel, Jean-Marc et Raymond Nadeau, *Statistique en gestion et en économie*, Éd. Gaétan Morin, 1988, p. 483-484.

donné sera d'autant plus grande que sa variation autour de la valeur la plus probable, soit le coefficient lui-même, est faible. C'est précisément ce que mesure l'erreur type du coefficient;

- **les *coefficients Beta standardisés* (*Standardized Coefficients*)**. Les coefficients non standardisés étant exprimés dans des unités de mesure différentes (dollars, prix au m², prix par année d'âge, etc.), ils ne peuvent être directement comparés. En les standardisant à l'aide de l'écart type de la variable à laquelle ils se rattachent, ils permettent de ranger les attributs résidentiels, selon l'importance de leur contribution respective à la formation des prix des immeubles vendus. Bien qu'ils ne puissent être utilisés à des fins prédictives, ils permettent, moyennant une manipulation relativement simple, de retracer *les coefficients d'élasticité* des variables explicatives du modèle, un concept cher aux économistes. Il s'agit, en fait, *d'une mesure de la sensibilité* de la valeur telle qu'établie par le modèle, relativement au changement à la marge dans la valeur de chacun des attributs;

- **le *Test de Student* (*t*)** est un indicateur *de la fiabilité statistique* des coefficients de régression. En référant à une table de Student, la valeur critique *« t »* correspondant au seuil de signification statistique de 0,05 (soit 0,025 de chaque côté de la distribution puisqu'il s'agit d'un test bilatéral) et à *N-1* degrés de liberté, est de 2,04 pour un échantillon excédant 31 observations et de 1,96 pour un échantillon qui tend vers l'infini (auquel cas la distribution de Student devient une distribution normale);

- **Les *VIF*s, ou *facteurs d'inflation de la variance***, constituent, à ce jour, l'indicateur le plus fiable de ce phénomène. En résumé, ils indiquent, dans quelle mesure, chaque variable indépendante du modèle est expliquée *par l'ensemble des autres variables explicatives.* Mesuré par l'équation suivante :

  VIF = (1/1- pouvoir explicatif des autres variables indépendantes),

  le VIF prendra la valeur *« 10 »* si la variable en question est expliquée par les autres dans une proportion de 90 %, *« 5 »* si le pouvoir explicatif est de 80 %, *« 2 »* s'il n'est que de 50 %, etc. Dans la littérature économétrique, les problèmes sérieux de multicolinéarité ne surviennent que si le VIF atteint ou excède la valeur *« 10 »*[1]. Dans les faits, toutefois, il importe de se montrer très vigilant dès que le VIF atteint la valeur *« 5 »*, les variables binaires, notamment, étant moins sensibles à cet indicateur que les variables métriques.

3.11.1.2 Exemple d'application

Pour les besoins de la démonstration de la première sous-étape de la procédure de régression standard, il est nécessaire de produire une première équation en introduisant l'ensemble des variables retenues dans la base de données, soit 39. Il s'agit là, bien sûr, d'un nombre excessif de descripteurs compte tenu de la taille de l'échantillon (N=132), mais qui permet de mettre en relief les problèmes potentiels de la méthode et les solutions à y apporter. Les résultats du tableau 3.5, présentés à la page suivante, amènent les commentaires ci-dessous.

[1] Voir à cet effet : Neter, J., Wasserman, W. and Kutner, M. H., *Applied Linear Statistical Models*, 2nd ed., 1985, R. D. Irwin, Homewood, IL, p. 391-392.

a) Sommaire du modèle

La première partie du tableau 3.5 représente les variables indépendantes utilisées pour définir le modèle à l'aide de la méthode « Enter ». La variable « VOISI_A » a été exclue par le modèle, puisque la limite de tolérance « 0 » a été atteinte dans le cas de cette variable.

La deuxième partie du tableau 3.5 présente le sommaire du modèle, lequel indique un « R^2 » ajusté de 0,725; en d'autres termes, les 38 variables indépendantes expliquent 72,5 % des variations du prix. Quant à l'erreur type de prévision, elle est de 6 280 $, soit 8,1 % du prix moyen de l'échantillon (77 253 $, voir le tableau 3.3), ce qui est déjà excellent.

b) Analyse de variance

La troisième partie du tableau 3.5 reproduit les résultats de *l'analyse de variance* permettant de vérifier si l'ensemble des variables indépendantes expliquent globalement « Y » de façon significative[2]. Comme il est possible de le constater, le *« F »* calculé est de 10,076, ce qui excède largement la valeur critique de 1,6. La probabilité que ce résultat soit dû au hasard est, en fait, inférieure à 0,0001 (colonne Sig). L'hypothèse nulle *« H_0 »* devra donc être exclue et, de ce fait, on peut conclure que l'ensemble des variables indépendantes du modèle explique les prix de vente de façon significative.

[2] Voir Martel et Nadeau, op. cit., p. 479-481.

Tableau 3.5 - Régression standard (mode « Enter »)
Variable dépendante = PRIXVTE
K (variables indépendantes) = 38
N = 132 observations

Modèle	Variables utilisées	Variables retirées	Méthode
1	MOIS, CUISEXTR, TERPENTE, CUISSUP, ASPIR, PROTECTI, ENTRÉESS, TERRIRR, PISCHT, LNAGEAPP, ABRI, SDBAIN, LNAIRHAB, FENRÉNOV, NBPLSTAT, THERMOPO, VOISI_B, LOCACOIN, PLAFCATH, PISCEXC, SDEAU, NBPLGAR, TOITRÉNO, GRGALERI, NBFOYER, SECTPROG, TTAMÉLOC, CHAUFFÉL, AIRFINSS, CLASSUP, SUPTER, PARDUR, AIREHAB, SDBSUP, SDBEXTRA, VOISI_C, AGEAPP, LNSUPTER[a]	,	Enter

a. Tolérance = ,000 limite à atteindre.

Sommaire du modèle

Modèle	R	R carré	R carré ajusté	Erreur type
1	,897[a]	,805	,725	$6,280

a. Prévision: (Constante), MOIS, CUISEXTR, TERPENTE, CUISSUP, ASPIR, PROTECTI, ENTRÉESS, TERRIRR, PISCHT, LNAGEAPP, ABRI, SDBAIN, LNAIRHAB, FENRÉNOV, NBPLSTAT, THERMOPO, VOISI_B, LOCACOIN, PLAFCATH, PISCEXC, SDEAU, NBPLGAR, TOITRÉNO, GRGALERI, NBFOYER, SECTPROG, TTAMÉLOC, CHAUFFÉL, AIRFINSS, CLASSUP, SUPTER, PARDUR, AIREHAB, SDBSUP, SDBEXTRA, VOISI_C, AGEAPP, LNSUPTER

Analyse de variance[b]

Modèle		Somme des carrés	df	moyenne des carrés	F	Sig.
1	Régression	1,510E+10	38	397365809	10,076	,000[a]
	Résidu	3667500832	93	39435492,8		
	Total	1,877E+10	131			

a. Prévision: (Constante), MOIS, CUISEXTR, TERPENTE, CUISSUP, ASPIR, PROTECTI, ENTRÉESS, TERRIRR, PISCHT, LNAGEAPP, ABRI, SDBAIN, LNAIRHAB, FENRÉNOV, NBPLSTAT, THERMOPO, VOISI_B, LOCACOIN, PLAFCATH, PISCEXC, SDEAU, NBPLGAR, TOITRÉNO, GRGALERI, NBFOYER, SECTPROG, TTAMÉLOC, CHAUFFÉL, AIRFINSS, CLASSUP, SUPTER, PARDUR, AIREHAB, SDBSUP, SDBEXTRA, VOISI_C, AGEAPP, LNSUPTER

b. Variable dépendante : PRIXVTE

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés			Statistiques de colinéarité	
		B	Erreur type	Beta	t	Sig.	Tolérance	VIF
1	(Constante)	$13,365	$37,182		.359	.720		
	VOISI_B	-$530	$1,599	-.021	-.331	.741	.526	1.90
	VOISI_C	-$7,134	$3,740	-.195	-1.907	.060	.200	4.99
	AGEAPP	-$802	$271	-.503	-2.964	.004	.073	13.70
	LNAGEAPP	$2,013	$3,202	.093	.629	.531	.097	10.35
	AIREHAB	$274	$64	.284	4.248	.000	.472	2.12
	LNAIRHAB	$1,365	$2,302	.034	.593	.555	.624	1.60
	AIRFINSS	$50	$24	.134	2.080	.040	.509	1.96
	CLASSUP	$4,168	$1,952	.137	2.135	.035	.510	1.96
	PARDUR	$569	$1,664	.024	.342	.733	.431	2.32
	FENRÉNOV	-$2,232	$1,613	-.081	-1.384	.170	.612	1.63
	TOITRÉNO	$1,047	$1,537	.040	.681	.498	.617	1.62
	SDBAIN	$5,099	$1,877	.213	2.717	.008	.342	2.92
	SDBSUP	-$1,057	$6,859	-.011	-.154	.878	.426	2.35
	SDBEXTRA	-$1,742	$1,615	-.078	-1.079	.283	.397	2.52
	SDEAU	$1,261	$2,277	.032	.554	.581	.649	1.54
	CUISSUP	-$4,007	$4,358	-.064	-.919	.360	.432	2.32
	CUISEXTR	$1,221	$868	.084	1.406	.163	.586	1.71
	CHAUFFÉL	$615	$1,668	.024	.369	.713	.489	2.04
	THERMOPO	$10,543	$2,849	.211	3.701	.000	.647	1.55
	ASPIR	$348	$1,405	.014	.248	.805	.698	1.43
	PROTECTI	$6,256	$4,444	.078	1.408	.163	.681	1.47
	PLAFCATH	$5,708	$3,186	.114	1.792	.076	.517	1.93
	ENTRÉESS	-$482	$1,656	-.017	-.291	.772	.593	1.69
	GRGALERI	-$63	$1,554	-.002	-.040	.968	.603	1.66
	NBFOYER	$1,639	$1,692	.061	.969	.335	.536	1.87
	NBPLGAR	$3,594	$1,084	.213	3.317	.001	.510	1.96
	ABRI	$577	$1,347	.024	.428	.669	.671	1.49
	NBPLSTAT	-$740	$469	-.092	-1.579	.118	.618	1.62
	PISCEXC	$9,927	$2,851	.220	3.482	.001	.525	1.91
	PISCHT	$4,644	$1,566	.173	2.965	.004	.614	1.63
	SUPTER	$5.36	$4.05	.202	1.325	.188	.091	11.03
	LNSUPTER	$2,506	$5,794	.075	.433	.666	.069	14.41
	TERRIRR	$163	$1,453	.007	.112	.911	.602	1.66
	TERPENTE	-$1,577	$2,019	-.044	-.781	.437	.653	1.53
	LOCACOIN	-$2,593	$1,632	-.090	-1.589	.116	.654	1.53
	TTAMÉLOC	$54	$1,802	.002	.030	.976	.405	2.47
	SECTPROG	$5,351	$3,347	.101	1.599	.113	.531	1.88
	MOIS	$247	$141	.117	1.751	.083	.467	2.14

a. Variable dépendante : PRIXVTE

c) Coefficients de régression

La quatrième et dernière portion du tableau 3.5, de la page précédente, porte sur les coefficients de régression individuels, soit les prix « implicites » proprement dits des attributs résidentiels.

Dans le cas présenté et selon l'analyse de variance (*df* = 131), il est possible de considérer que si la valeur *« t »* (*Test de Student*) calculée atteint ou excède 2,00 (en valeur absolue), le coefficient de régression est statistiquement significatif et, par conséquent, différent de zéro. La colonne (Sig), étant un corollaire de la précédente, permet d'établir de façon précise le degré de signification statistique de chacun des paramètres de régression.

d) Appréciation des variables explicatives utilisées

L'analyse du tableau 3.5 amène à conclure qu'à ce stade de l'analyse, relativement peu de coefficients sont effectivement significatifs au seuil 0,05. C'est, cependant, le cas des paramètres relatifs aux variables « AGEAPP », « AIREHAB », « AIRFINISS », « CLASSUP », « SDBAIN », « THERMOPO », « NBPLGAR », « PISCEXC » et « PISCHT », dont c'est possible, d'ores et déjà, de supposer qu'ils constituent, pour le segment de marché considéré, des déterminants importants de la valeur marchande des immeubles. En raison, toutefois, de la très forte multicolinéarité affectant plusieurs variables du modèle, il est inapproprié, à ce stade-ci, de se prononcer sur les prix modélisés des caractéristiques.

La dernière colonne du tableau fournit précisément une mesure de multicolinéarité. Comme il est possible de le constater, plusieurs variables ou séries de variables affichent des VIFs excédant largement « 10 ». C'est le cas, en particulier, du groupe « AGEAPP » et « LNAGEAP », deux attributs fortement reliés et qu'il ne convient pas, évidemment, de retrouver dans une même équation de régression, ainsi que des variables « SUPTER » et « LNSUPTER ». Les conséquences observées sont classiques, soit des coefficients dont les tests *« t »* et leur probabilité respective sont invalidés et dont le signe, du moins dans le cas du logarithme de l'âge apparent et de la superficie de terrain, est contraire à la logique.

3.11.2 RÉDUCTION DU NOMBRE DE VARIABLES

À la suite de cette première tentative, l'analyste cherche à réduire substantiellement le nombre de variables explicatives, de façon à stabiliser les coefficients de régression et à en arriver éventuellement à un modèle qui soit à la fois performant et robuste. L'analyste retire de l'équation les caractéristiques ne répondant pas aux critères de signification statistique retenus (seuil de 0,05) par une procédure itérative « manuelle », de type essais et erreurs.

Il est important de souligner, ici, qu'une réduction du « R^2 » et une hausse de l'erreur type de prévision (SEE) sont possibles chaque fois qu'une variable est retirée du modèle. Toutefois, dans la mesure où les attributs éliminés n'ont qu'un faible pouvoir explicatif et sont, par ailleurs, susceptibles de générer de la multicolinéarité, leur retrait demeure globalement bénéfique puisque la détérioration des performances explicative et prédictive de l'équation de régression n'est que très légère, alors que sa robustesse statistique s'en trouve sensiblement haussée, du fait de l'augmentation des degrés de liberté qui en découle. Ceci se traduit par un test *« F »* plus fort.

3.11.2.1 Exemple d'application

Par cette procédure, le nombre de descripteurs passe progressivement de 38 à 16 descripteurs, puis à 13 et, enfin, à 10 variables explicatives. C'est l'objet des trois *tableaux suivants, 3.6.a, 3.6.b et 3.6.c,* comportant trois sections.

a) Première phase d'épuration du modèle de régression

Dans la première phase d'épuration du modèle de régression des variables problématiques, présenté au tableau 3.6.a, ne sont retenus que les coefficients dont le degré de signification statistique est de 0,15 ou moins. Ce qui a permis de retrancher 23 descripteurs des 38 initialement retenus. L'analyse des résultats nous indique que le « R^2 » ajusté s'est amélioré (0,738 contre 0,725 antérieurement, voir le tableau 3.5) et que l'erreur de prévision a diminué (6 128 $ contre 6 280 $). Mais c'est surtout au chapitre de la valeur *« F »* que l'amélioration est la plus marquée, cette dernière passant de 10,08 à 24,05. Quant aux coefficients de régression, ils sont tous significatifs au seuil de 0,05 (voir colonne *« t »*), à l'exception de ceux des variables « PROTECTI », « SECTPROG » et « MOIS », qui seront donc retirées de l'équation. Enfin, la multicolinéarité, telle qu'indiquée par les VIFs, a, à toutes fins pratiques, disparu du modèle.

Tableau 3.6.a - Réduction du nombre de variables indépendantes de 38 à 16

Sommaire du modèle

Modèle	R	R carré	R carré ajusté	Erreur type
1	,877[a]	,770	,738	$6,128

a. Prévision: (Constante), VOISI_C, CLASSUP, NBPLSTAT, PISCHT, PROTECTI, SDBAIN, THERMOPO, NBPLGAR, SECTPROG, AIREHAB, PLAFCATH, MOIS, PISCEXC, AIRFINSS, AGEAPP, SUPTER

Analyse de variance[b]

Modèle		Somme des carrés	df	Moyenne des carrés	F	Sig.
1	Régression	1,44E+10	16	903046179,5	24,047	,000[a]
	Residu	4,32E+09	115	37553588,70		
	Total	1,88E+10	131			

a. Prévision : (Constante), VOISI_C, CLASSUP, NBPLSTAT, PISCHT, PROTECTI, SDBAIN, THERMOPO, NBPLGAR, SECTPROG, AIREHAB, PLAFCATH, MOIS, PISCEXC, AIRFINSS, AGEAPP, SUPTER

b. Variable dépendante : PRIXVTE

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés			Statistiques de colinéarité	
		B	*Erreur type*	*Beta*	*t*	*Sig.*	*Tolérance*	*VIF*
1	*(Constante)*	**$37,372**	**$5,966**		**6.264**	**.000**		
	PROTECTI	**$4,647**	**$3,808**	**.058**	**1.220**	**.225**	**.883**	**1.13**
	AIRFINSS	**$58**	**$20**	**.157**	**2.911**	**.004**	**.686**	**1.46**
	PISCHT	**$3,961**	**$1,292**	**.148**	**3.066**	**.003**	**.859**	**1.16**
	NBPLGAR	**$3,567**	**$868**	**.211**	**4.108**	**.000**	**.756**	**1.32**
	SECTPROG	**$4,817**	**$2,745**	**.091**	**1.755**	**.082**	**.752**	**1.33**
	PLAFCATH	**$6,098**	**$2,758**	**.122**	**2.211**	**.029**	**.657**	**1.52**
	THERMOPO	**$7,956**	**$2,444**	**.159**	**3.256**	**.001**	**.837**	**1.20**
	MOIS	**$206**	**$114**	**.098**	**1.806**	**.073**	**.681**	**1.47**
	PISCEXC	**$10,294**	**$2,435**	**.228**	**4.227**	**.000**	**.685**	**1.46**
	NBPLSTAT	**-$863**	**$402**	**-.107**	**-2.145**	**.034**	**.800**	**1.25**
	SUPTER	**$6.50**	**$1.74**	**.245**	**3.738**	**.000**	**.467**	**2.14**
	AIREHAB	**$325**	**$51**	**.337**	**6.367**	**.000**	**.714**	**1.40**
	SDBAIN	**$2,988**	**$1,276**	**.125**	**2.341**	**.021**	**.705**	**1.42**
	AGEAPP	**-$642**	**$93**	**-.403**	**-6.929**	**.000**	**.593**	**1.69**
	CLASSUP	**$4,892**	**$1,762**	**.161**	**2.776**	**.006**	**.597**	**1.68**
	VOISI_C	**-$6,330**	**$2,745**	**-.173**	**-2.306**	**.023**	**.354**	**2.82**

a. Variable dépendante: PRIXVTE

b) Deuxième phase d'épuration du modèle de régression

Avec le retrait des trois variables dont les coefficients de régression sont non significatifs au seuil de 0,05 (il s'agit ici des variables « PROTECTI », « SECTPROG » et « MOIS »), le modèle présenté au tableau 3.6.b subit une légère baisse de ses performances explicative et prédictive; le « R^2 » ajusté chutant à 0,728, alors que l'erreur type d'estimation augmente à 6 237 $. Par contre, une amélioration sensible du test *« F »* (28,04) est observée. Quant à l'analyse des valeurs *« t »*, elle indique que si la force statistique de plusieurs coefficients de régression s'en trouve réduite ou, au mieux, maintenue, il n'en va pas de même des principaux descripteurs (« AGEAPP », « AIREHAB », « CLASSUP ») dont le test de Student en ressort renforcé.

Tableau 3.6.b - Réduction de 16 à 13 variables indépendantes

Sommaire du modèle

Modèle	*R*	*R carré*	*R carré ajusté*	*Erreur type*
1	**,869[a]**	**,755**	**,728**	**$6,237**

a. Prévision : (Constante), VOISI_C, CLASSUP, NBPLSTAT, PISCHT, AIREHAB, SDBAIN, THERMOPO, NBPLGAR, PLAFCATH, AIRFINSS, PISCEXC, AGEAPP, SUPTER

Analyse de variance[b]

Modèle		Somme des carrés	df	Moyenne des carrés	F	Sig.
1	Régression	1,418E+10	13	1,09E+09	28,036	,000[a]
	Résidu	4,590E+09	118	38898549		
	Total	1,877E+10	131			

a. Prévision : (Constante), VOISI_C, CLASSUP, NBPLSTAT, PISCHT, AIREHAB, SDBAIN, THERMOPO, NBPLGAR, PLAFCATH, AIRFINSS, PISCEXC, AGEAPP, SUPTER

b. Variable dépendante : PRIXVTE

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.	Statistiques de colinéarité	
		B	Erreur type	Beta			Tolérance	VIF
1	(Constant)	$40,700	$5,308		7.667	.000		
	AIRFINSS	$52	$20	.139	2.596	.011	.724	1.38
	PISCHT	$3,372	$1,285	.126	2.623	.010	.899	1.11
	NBPLGAR	$3,485	$880	.207	3.961	.000	.763	1.31
	PLAFCATH	$4,385	$2,723	.088	1.610	.110	.698	1.43
	THERMOPO	$6,410	$2,414	.128	2.655	.009	.888	1.13
	PISCEXC	$10,051	$2,448	.223	4.105	.000	.702	1.42
	NBPLSTAT	-$568	$393	-.071	-1.448	.150	.869	1.15
	SUPTER	$5.71	$1.72	.215	3.324	.001	.496	2.02
	AIREHAB	$349	$50	.362	6.939	.000	.763	1.31
	SDBAIN	$2,938	$1,264	.123	2.324	.022	.744	1.34
	AGEAPP	-$673	$93	-.422	-7.264	.000	.614	1.63
	CLASSUP	$5,319	$1,756	.175	3.028	.003	.622	1.61
	VOISI_C	-$3,463	$2,450	-.095	-1.413	.160	.461	2.17

a. Variable dépendante: PRIXVTE

c) Troisième phase d'épuration du modèle de régression

Au cours de la troisième et dernière phase d'épuration (tableau 3.6.c), les variables, dont le test *« t »* était inférieur à *« 2 »,* lors de la phase précédente, sont retranchées de l'analyse, c.-à-d. celles qui ne rencontraient pas le seuil de signification de 0,05. Il s'agit ici des variables « VOISI_C », « PLAFCATH » et « NBPLSTAT ». Cette fois, la hausse marginale de la valeur *« F »,* passant de 28,04 à 34,82, est encore plus substantielle que dans le cas précédent et compense largement pour la perte des pouvoirs explicatif et prédictif du modèle de régression. En outre, tous les paramètres de régression sont significatifs au seuil de 0,05.

Tableau 3.6.c - Réduction de 13 à 10 variables indépendantes

Sommaire du modèle

Modèle	R	R carré	R carré ajusté	Erreur type
1	,861[a]	,742	,721	$6,325

a. Prévision : (Constante), CLASSUP, NBPLGAR, PISCHT, THERMOPO, SDBAIN, SUPTER, AIREHAB, AIRFINSS, PISCEXC, AGEAPP

Analyse de variance[b]

Modèle		Somme des carrés	df	Moyenne des carrés	F	Sig.
1	Régression	1,39E+10	10	1392717352	34,816	,000[a]
	Résidu	4,84E+09	121	40001884,71		
	Total	1,88E+10	131			

a. Prévision : (Constante), CLASSUP, NBPLGAR, PISCHT, THERMOPO, SDBAIN, SUPTER, AIREHAB, AIRFINSS, PISCEXC, AGEAPP

b. Variable dépendante : PRIXVTE

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.	Statistiques de colinéarité	
		B	Erreur type	Beta			Tolérance	VIF
1	(Constante)	$38,244	$5,241		7.297	.000		
	AIRFINSS	$46	$20	.124	2.356	.020	.768	1.30
	PISCHT	$2,997	$1,288	.112	2.326	.022	.920	1.09
	NBPLGAR	$3,470	$874	.206	3.971	.000	.795	1.26
	THERMOPO	$5,995	$2,426	.120	2.471	.015	.904	1.11
	PISCEXC	$9,467	$2,427	.210	3.901	.000	.735	1.36
	SUPTER	$3.68	$1.36	.139	2.697	.008	.808	1.24
	AIREHAB	$367	$50	.380	7.286	.000	.783	1.28
	SDBAIN	$3,294	$1,265	.138	2.605	.010	.764	1.31
	AGEAPP	-$651	$86	-.408	-7.541	.000	.728	1.37
	CLASSUP	$6,617	$1,654	.217	4.002	.000	.722	1.39

a. Variable dépendante : PRIXVTE

3.11.3 TRANSFORMATION MATHÉMATIQUE DES VARIABLES INDÉPENDANTES

Une procédure « automatique », visant à éliminer de l'équation les attributs ne contribuant pas, de façon significative, à l'explication des prix et en arriver à la spécification d'un modèle optimal, est présentée plus loin. Auparavant, toutefois, il convient de vérifier s'il est possible d'améliorer les performances de l'équation précédente en transformant mathématiquement certaines variables[1].

3.11.3.1 Exemple d'application

Dans un premier temps, il convient d'opérer sur les variables « AGEAPP » (âge apparent), « AIREHAB » (aire habitable) et « SUPTER » (superficie de terrain) une transformation logarithmique. Les résultats de l'analyse font l'objet du *tableau 3.7*. Comme il est possible de le constater, cette tentative se solde par une détérioration prononcée du « R^2 » ajusté, de l'erreur type de prévision et du test *« F »*, ce qui suggère que la transformation logarithmique n'est pas appropriée ici, du moins sur deux des trois variables identifiées, soit l'âge apparent et l'aire habitable, dont la stabilité des paramètres s'en trouve fortement affectée. Cette transformation a, cependant, un effet bénéfique sur le comportement de la variable « SUPTER » dont l'erreur type du coefficient diminue sensiblement, faisant ainsi passer la valeur *« t »* de 2,70 à 3,33. Un tel résultat n'est pas surprenant puisque, comme l'étape 4 a permis de le constater, la distribution des superficies de terrain est fortement étalée vers la droite. L'utilisation de données logarithmiques a pour effet d'atténuer cet étalement et de rendre plus efficace le recours à la régression linéaire multiple. Dans l'exemple utilisé ici, cette transformation ne s'avère pas indispensable à la performance globale du modèle de régression et relève plutôt du choix de l'analyste que d'une nécessité absolue.

Tableau 3.7 - Transformation logarithmique de certaines variables indépendantes

Sommaire du modèle

Modèle	*R*	*R carré*	*R carré ajusté*	*Erreur type*
1	**,795[a]**	**,631**	**,601**	**$7,560**

a. Prévision : (Constante), LNAGEAPP, PISCEXC, THERMOPO, LNAIRHAB, LNSUPTER, PISCHT, AIRFINSS, NBPLGAR, SDBAIN, CLASSUP

Analyse de variance[b]

Modèle		*Somme des carrés*	*df*	*Moyenne des carrés*	*F*	*Sig.*
1	*Régression*	**1,185E+10**	**10**	**1,19E+09**	**20,733**	**,000[a]**
	Résidu	**6,916E+09**	**121**	**57160518**		
	Total	**1,877E+10**	**131**			

a. Prévision : (Constante), LNAGEAPP, PISCEXC, THERMOPO, LNAIRHAB, LNSUPTER, PISCHT, AIRFINSS, NBPLGAR, SDBAIN, CLASSUP

b. Variable dépendante : PRIXVTE

[1] Le logiciel SPSS offre à l'analyste un moyen efficace d'établir la relation fonctionnelle optimale entre la variable dépendante et une variable indépendante.

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés			Statistiques de colinéarité	
		B	Erreur type	Beta	t	Sig.	Tolérance	VIF
1	(Constante)	$8,829	$18,097		.488	.627		
	LNAIRHAB	$5,794	$2,239	.146	2.588	.011	.957	1.04
	AIRFINSS	$56	$23	.150	2.373	.019	.766	1.31
	PISCHT	$3,340	$1,538	.125	2.172	.032	.923	1.08
	LNSUPTER	$6,963	$2,093	.209	3.327	.001	.771	1.30
	THERMOPO	$9,204	$2,843	.184	3.237	.002	.941	1.06
	CLASSUP	$9,011	$1,957	.296	4.605	.000	.737	1.36
	NBPLGAR	$3,726	$1,051	.221	3.546	.001	.785	1.27
	PISCEXC	$10,105	$2,823	.224	3.579	.000	.776	1.29
	SDBAIN	$3,859	$1,517	.161	2.544	.012	.759	1.32
	LNAGEAPP	-$6,631	$1,368	-.305	-4.848	.000	.768	1.30

a. Variable dépendante : PRIXVTE

3.11.4 TRANSFORMATION MATHÉMATIQUE DE LA VARIABLE DÉPENDANTE

L'application d'une transformation logarithmique sur la variable dépendante (soit les prix de vente) a pour effet de générer un modèle *de type multiplicatif* dans lequel les termes de l'équation se multiplient les uns les autres plutôt que de suivre une logique additive (voir le point 3.9.2 de ce chapitre). Les coefficients de régression découlant de l'application de cette forme semi-logarithmique (dénommée *semi-log*) se présentent sous la forme de termes d'ajustement positifs ou négatifs qui, moyennant une simple transformation mathématique, deviennent *des facteurs d'ajustement* supérieurs ou inférieurs à l'unité. Il s'agit, pour cela, d'en calculer *l'antilogarithme*, en appliquant, aux paramètres de la régression, une transformation exponentielle. Ici, les contributions marginales des attributs résidentiels sont donc cumulatives, l'ordonnée à l'origine (le terme constant de l'équation) constituant la valeur de base.

3.11.4.1 Exemple d'application

Le *tableau 3.8,* de la page suivante, reproduit une transformation logarithmique du prix de vente. Le modèle ainsi obtenu génère des performances se comparant à celles du tableau 3.6.c, avec toutefois un « R^2 » ajusté légèrement inférieur (0,71 versus 0,72) et une erreur de prévision légèrement supérieure (8,6 % versus 8,2 % du prix moyen)[1]. On note par contre une amélioration de la force statistique des coefficients de régression, qui sont tous significatifs au seuil de 0,01. Quant à l'interprétation des paramètres de l'équation, elle est relativement aisée. En leur appliquant la transformation suggérée plus haut, il est possible de conclure, par exemple, qu'une unité d'évaluation, dont l'état est qualifié de *supérieur,* vaut 7,9 % de plus (facteur d'ajustement de 1,0786) qu'une autre unité de qualité standard. *Une piscine excavée* rajoute, pour sa part, 11,7 % (facteur d'ajustement de 1,1166) à la valeur marchande de l'unité d'évaluation.

[1] Dans le premier cas, le coefficient de variation *(CV),* soit l'erreur type relative, s'obtient en divisant l'erreur type d'estimation (6 325 $) par le prix moyen de l'échantillon (77 253 $). Dans le second cas, considérer l'antilog de l'erreur type (0,08264), ce qui donne une erreur en pourcentage de 8,6 %.

Tableau 3.8 - Transformation logarithmique sur la variable dépendante

Sommaire du modèle

Modèle	R	R Carré	R Carré ajusté	Erreur type
1	,856[a]	,732	,710	.08264

a. Prévision : (Constante), SUPTER, CLASSUP, PISCHT, THERMOPO, SDBAIN, AIREHAB, NBPLGAR, AIRFINSS, PISCEXC, AGEAPP

Analyse de variance[b]

Modèle		Somme des carrés	df	Moyenne des carrés	F	Sig.
1	Régression	2,260	10	,226	33,087	,000[a]
	Résidu	,826	121	6,830E-03		
	Total	3,086	131			

a. Prévision : (Constante), SUPTER, CLASSUP, PISCHT, THERMOPO, SDBAIN, AIREHAB, NBPLGAR, AIRFINSS, PISCEXC, AGEAPP

b. Variable dépendante : LNPRIXVT

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés			Statistiques de colinéarité	
		B	Erreur type	Beta	t	Sig.	Tolérance	VIF
1	(Constante)	10.75316	,06849		157.011	.000		
	AIRFINSS	.00071	,00026	.148	2.764	.007	.768	1.30
	PISCHT	.04379	,01683	.128	2.601	.010	.920	1.09
	THERMOPO	.08760	,03170	.137	2.763	.007	.904	1.11
	CLASSUP	.07563	,02161	.194	3.500	.001	.722	1.39
	NBPLGAR	.04439	,01142	.205	3.888	.000	.795	1.26
	PISCEXC	.11031	,03171	.191	3.479	.001	.735	1.36
	SDBAIN	.04564	,01653	.149	2.762	.007	.764	1.31
	AIREHAB	.00455	,00066	.367	6.911	.000	.783	1.28
	AGEAPP	-.00863	,00113	-.422	-7.651	.000	.728	1.37
	SUPTER	.00005	,00002	.140	2.682	.008	.808	1.24

a. Variable dépendante : LNPRIXVT

3.11.5 PROCÉDURE DE RÉGRESSION PAR ÉTAPE

L'étape 6 se termine par la présentation d'une procédure de régression faisant le bonheur de tout analyste, puisqu'elle permet de reproduire « automatiquement » le processus d'épuration décrit précédemment et d'en arriver à la forme optimale du modèle, soit celle maximisant les performances avec le minimum de variables explicatives. Il s'agit de la procédure de *régression par étape* ou *Stepwise Regression*. En vertu de cette procédure, les variables indépendantes, soumises en bloc par l'analyste, sont introduites une à une dans l'équation de régression, en fonction de leur contribution marginale à l'explication du phénomène (« R^2 » partiel)[1] et de leur degré de signification statistique (par application d'un test *« F »* individuel). Chaque étape de la procédure aboutit à l'ajout d'une nouvelle variable, le nombre total d'étapes exécutées correspondant au nombre de variables explicatives qui seront finalement retenues. Chaque attribut ajouté apporte ainsi une contribution marginale de moins en moins importante, bien que toujours significative, à l'explication des prix, la procédure s'arrêtant lorsque l'ajout d'une variable, dont le pouvoir explicatif est trop faible, se traduit par une détérioration de la performance globale du modèle.

Bien que la procédure de régression par étape ne puisse se substituer au jugement de l'analyste, demeurant libre de « forcer » dans l'équation certaines caractéristiques dont il tient à mesurer la contribution, elle demeure un outil d'aide à la décision extrêmement puissant et constitue, d'une façon générale, le meilleur remède contre la multicolinéarité excessive.

3.11.5.1 Exemple d'application

Le *tableau 3.9,* de la page suivante, présente les résultats d'une application de la procédure de régression par étape pour l'échantillon de 132 immeubles vendus; les 38 variables originelles sont introduites dans l'analyse.

La première partie du tableau 3.9 (variables utilisées/retirées) reproduit l'ordre d'entrée des variables dans l'équation de régression. Il est intéressant de constater que, des onze variables retenues, c'est la caractéristique « CLASSUP » qui s'impose comme la plus importante, suivie des variables « AIREHAB » et « AGEAPP ». Leur contribution marginale cumulative au « R^2 » se retrouve dans la seconde partie du tableau (sommaire du modèle), avec l'erreur type d'estimation obtenue à chaque étape de la procédure. Ainsi, les trois premières variables expliquent à elles seules 58,1% (« R^2 » ajusté) de la variation des prix de vente, alors que les onze descripteurs génèrent une performance explicative de 73,2 %. Quant à l'erreur type d'estimation, elle passe de 10 180 $ (13,2 % du prix moyen) à l'étape *n° 1* à 6 198 $ (8,0 %) à l'étape *n° 11*.

[1] Le « R^2 » partiel estime l'importance de la contribution marginale d'une caractéristique donnée, compte tenu des attributs déjà inclus dans le modèle (voir Martel et Nadeau, op. cit. p. 484-485).

Tableau 3.9 - Application de la procédure de régression par étape

K (variables indépendantes) : Départ = 38; Arrivée = 11
N = 132 observations

Variables utilisées/Retirées[a]

Modèle	Variables utilisées	Variables retirées	Méthodes
1	CLASSUP	,	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
2	AIREHAB	,	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
3	AGEAPP	,	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
4	PISCEXC	,	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
5	NBPLGAR	,	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
6	AIRFINSS	,	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
7	SDBAIN	,	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
8	SUPTER	,	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
9	THERMOPO	,	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
10	PISCHT	,	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
11	CUISEXTR	,	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).

a. Variable dépendante : PRIXVTE

Sommaire du modèle

Modèle	R	R carré	R carré ajusté	Erreur type
1	,531	,282	,277	$10,180
2	,680	,462	,454	$8,844
3	,769	,591	,581	$7,745
4	,792	,627	,615	$7,427
5	,820	,672	,659	$6,992
6	,835	,698	,683	$6,736
7	,842	,708	,692	$6,646
8	,849	,721	,703	$6,526
9	,855	,731	,711	$6,438
10	,861	,742	,721	$6,325
11	,869	,754	,732	$6,198

Analyse de variance

Modèle		Somme des carrés	df	Moyenne des carrés	F	Sig.
1	Régression	5,30E+09	1	5,2954E+09	51,098	,000
	Résidu	1,35E+10	130	1,0363E+08		
	Total	1,88E+10	131			
2	Régression	8,68E+09	2	4,3393E+09	55,484	,000
	Résidu	1,01E+10	129	7,8208E+07		
	Total	1,88E+10	131			
3	Régression	1,11E+10	3	3,6966E+09	61,630	,000
	Résidu	7,68E+09	128	5,9981E+07		
	Total	1,88E+10	131			
4	Régression	1,18E+10	4	2,9404E+09	53,302	,000
	Résidu	7,01E+09	127	5,5164E+07		
	Total	1,88E+10	131			
5	Régression	1,26E+10	5	2,5217E+09	51,587	,000
	Résidu	6,16E+09	126	4,8881E+07		
	Total	1,88E+10	131			
6	Régression	1,31E+10	6	2,1825E+09	48,093	,000
	Résidu	5,67E+09	125	4,5380E+07		
	Total	1,88E+10	131			
7	Régression	1,33E+10	7	1,8986E+09	42,983	,000
	Résidu	5,48E+09	124	4,4171E+07		
	Total	1,88E+10	131			
8	Régression	1,35E+10	8	1,6911E+09	39,706	,000
	Résidu	5,24E+09	123	4,2590E+07		
	Total	1,88E+10	131			
9	Régression	1,37E+10	9	1,5234E+09	36,755	,000
	Résidu	5,06E+09	122	4,1448E+07		
	Total	1,88E+10	131			
10	Régression	1,39E+10	10	1,3927E+09	34,816	,000
	Résidu	4,84E+09	121	4,0002E+07		
	Total	1,88E+10	131			
11	Régression	1,42E+10	11	1,2870E+09	33,497	,000
	Résidu	4,61E+09	120	3,8421E+07		
	Total	1,88E+10	131			

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.	Statistiques de colinéarité	
		B	Erreur type	Beta			Tolérance	VIF
11	(Constant)	$39,810	$5,176		7.691	.000		
	CLASSUP	$6,075	$1,636	.200	3.715	.000	.709	1.411
	AIREHAB	$339	$51	.352	6.703	.000	.744	1.344
	AGEAPP	-$654	$85	-.410	-7.726	.000	.728	1.373
	PISCEXC	$9,535	$2,379	.212	4.009	.000	.735	1.361
	NBPLGAR	$3,149	$866	.187	3.634	.000	.776	1.288
	AIRFINSS	$41	$19	.111	2.134	.035	.759	1.317
	SDBAIN	$3,357	$1,240	.140	2.708	.008	.764	1.309
	SUPTER	$3.60	$1.34	.135	2.688	.008	.807	1.239
	THERMOPO	$6,592	$2,390	.132	2.758	.007	.895	1.118
	PISCHT	$3,415	$1,274	.128	2.680	.008	.904	1.106
	CUISEXTR	$1,769	$724	.122	2.445	.016	.822	1.217

a. Variable dépendante : PRIXVTE

À la troisième partie du tableau 3.9 de la page précédente (analyse de variance), la valeur *« F »* atteint un maximum de 61,63 au modèle *n° 3* pour se stabiliser à 33,50 à la fin du processus. Tel que démontré, la procédure *Stepwise* produit, au chapitre des performances globales, les meilleurs résultats obtenus jusqu'ici.

L'analyse de la quatrième et dernière partie du tableau 3.9 (Coefficients), ne reproduisant, en l'occurrence, que les résultats du dernier modèle, démontre en outre que tous les coefficients de régression, s'avérant cohérents tant par leur amplitude que par leur signe, sont significatifs au seuil 0,01, à l'exception de celui de la variable « AIRFINISS » qui l'est au seuil 0,05. Une comparaison avec les résultats obtenus en utilisant l'approche « manuelle » (tableau 3.6.c) indique par ailleurs que la plupart des paramètres générés avec la régression par étape sont statistiquement plus solides.

Il importe de retenir enfin, quelle que soit l'approche utilisée, qu'aucune des variables de secteur initialement introduites dans l'analyse ne s'avère statistiquement significative. Cela implique donc qu'il n'existe pas de différence de prix significative entre les résidences des trois secteurs de la ville sous étude, puisque ces différences sont déjà prises en compte par les autres variables explicatives de l'équation.

Bien que ce modèle soit très performant, en particulier au plan prédictif, il est possible de l'améliorer par l'analyse et l'élimination des résidus extrêmes qui biaisent l'interprétation des coefficients de régression. C'est l'objet de l'*étape 7* présentée aux pages suivantes.

3.12 ÉTAPE 7 : ANALYSE DES RÉSIDUS

Comme il est mentionné au point 3.4.3 de ce chapitre, l'analyse de régression multiple génère une *estimation* de la véritable relation entre les prix de vente et les caractéristiques résidentielles déterminant la valeur réelle des unités d'évaluation. Cela implique, notamment, que pour chaque prix de vente observé, il existe une valeur estimée, la différence entre les deux constituant un résidu. En analyse de régression, le problème réside non pas dans l'existence de résidus, qui est inévitable, mais dans l'existence de résidus *extrêmes* ayant pour effet de fausser les paramètres de la régression. Ces résidus extrêmes proviennent de diverses sources, dont la mauvaise spécification des variables et de la forme fonctionnelle du modèle, l'omission de caractéristiques majeures et l'information imparfaite dont dispose l'évaluateur sur les attributs des immeubles vendus. À titre d'exemple, certaines propriétés « haut de gamme » présentent des caractéristiques uniques qui, bien que prises en compte par le marché, ne peuvent être modélisées en raison de leur trop faible fréquence. En outre, plusieurs attributs du voisinage telles la qualité de la vue, la proximité d'une autoroute, etc., sont autant de dimensions importantes que l'analyste omet souvent d'intégrer au modèle, faute de pouvoir les mesurer adéquatement. Enfin, toutes les circonstances et conditions d'une transaction n'étant pas nécessairement portées au contrat de vente, il est possible que certaines particularités affectant le prix payé échappent à l'analyste.

Ces informations manquantes auront un effet d'autant plus nocif sur la valeur des coefficients de régression, que l'écart entre les prix observés et les prix estimés par l'équation sera prononcé. Il est donc justifié de retirer ces observations de l'analyse si cet écart, c.-à-d. le résidu, excède une valeur « raisonnable ».

3.12.1 IDENTIFICATION DES RÉSIDUS DÉLINQUANTS

Pour établir la valeur « raisonnable » des résidus, il est important d'identifier *les résidus standardisés*[1] du modèle. Tout résidu supérieur à *2 ou 3 SEE* , selon le jugement de l'analyste, est ainsi retiré de l'analyse. Dans la mesure où les résidus du modèle de régression sont distribués normalement, un tel retrait implique que la taille de l'échantillon se trouve réduite d'environ 5 %, tout au plus. Cette opération ne doit se faire qu'une seule fois, soit lors de la mise au point finale du modèle. Toutefois, l'analyste doit, au préalable, et dans la mesure du possible, *justifier le retrait des ventes problématiques* par une analyse approfondie des facteurs causant ces résidus extrêmes.

3.12.1.1 Exemple d'application

Les résultats de l'analyse des résidus du modèle de régression apparaissent au *tableau 3.10* de la page suivante*.* Une première partie *(analyse des résidus)* présente les observations ayant des résidus standardisés supérieurs à *2 SEE*, leur prix de vente, les valeurs prédites et les résidus absolus sont exprimés en dollars. La seconde partie du tableau fournit les principales statistiques relatives à ces résidus. C'est l'objet de l'analyse suivante, alors que quatre transactions générant des résidus extrêmes sont relevées : Il s'agit des ventes n^{os} *119*, *135*, *146* et *151*. Alors que le modèle sous-évalue les deux dernières propriétés (résidus positifs de 14 274 $ et 12 606 $ respectivement), il surévalue les deux premières (résidus négatifs de -15 199 $ et -13 816 $ respectivement).

[1] Les *résidus standardisés* du modèle sont obtenus en divisant les résidus exprimés en dollars (dans la forme linéaire) par l'erreur type d'estimation (SEE).

Tableau 3.10 - Analyse des résidus du modèle de régression par étape

K (variables indépendantes) = 11
N = 132 observations
Variable dépendante = PRIXVTE

Analyse des résidus

Numéro de la vente	Résidu standardisé	PRIXVTE	Valeur prédite	Résidu absolu
119	2,303	$75,500	$61,226	$14,274
135	2,034	$84,000	$71,394	$12,606
146	-2,452	$67,200	$82,399	-$15,199
151	-2,229	$74,000	$87,816	-$13,816

Statistiques sur les résidus

	Minimum	Maximum	Moyenne	Écart type	Nombre
Valeur prédite	$58,914	$106,848	$77,253	$10,396	132
Résidu	-$15,199	$14,274	$.00	$5,933	132
Valeur standardisée prédite	-1,764	2,847	,000	1,000	132
Résidu standardisé	-2,452	2,303	,000	,957	132

Une analyse plus fouillée de ces transactions révèle des éléments contribuant à expliquer les résidus obtenus :

N° de vente	Adresse civique	Prix de vente/Évaluation/ Valeur estimée	Commentaires
119	35, Assomption	Prix de vente :75 500 $ Évaluation : 61 900 $ Val. estimée : 61 226 $	• Il s'agit d'une vente entre parents, l'un des deux acheteurs étant la fille du vendeur. • Diverses considérations financières ont été prises par les parties.
135	14, Des Plaines	Prix de vente :84 000 $ Évaluation : 76 400 $ Val. estimée : 71 394 $	• Cette propriété a été achetée par une entreprise pour sa localisation. • Cet élément n'est pas pris en compte par le modèle.
146	6, rue Jeanne	Prix de vente :67 200 $ Évaluation : 77 700 $ Val. estimée : 82 399 $	• Il s'agit d'une vente de succession, réalisée à rabais par des vendeurs résidant à l'extérieur de la région; la propriété est, par ailleurs, située dans un secteur périphérique dénué de services d'aqueduc et d'égout. • Aucun élément n'est pris en compte par le modèle.
151	6, rue Marie-Ève	Prix de vente :74 000 $ Évaluation : 81 300 $ Val. estimée : 87 816 $	• la propriété est située dans un secteur périphérique dénué de services d'aqueduc et d'égout. • Aucun élément n'est pris en compte par le modèle.

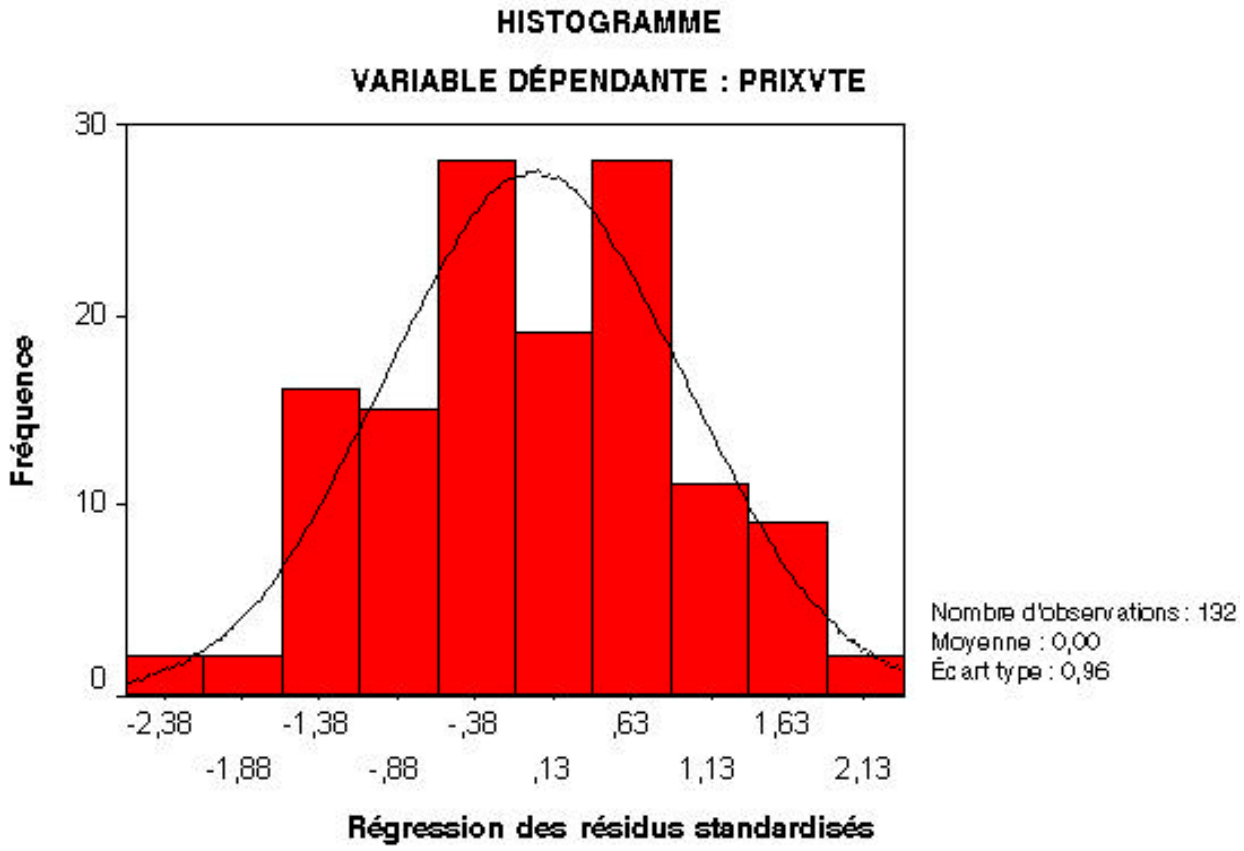
L'écart prononcé, entre le prix observé pour ces immeubles vendus et la valeur estimée par le modèle, s'explique fort bien dans tous les cas. En ce qui a trait aux ventes *n^os^ 135* et *151*, la cause de l'erreur réside dans l'omission d'attributs propres au site ou de caractéristiques de voisinage. Quant aux ventes *n^os^ 119* et *146*, leur prix reflète des conditions très particulières faisant qu'elles ne peuvent être considérées comme des ventes *bona fide* au sens strict du terme. Pour cette raison, il est préférable de les exclure de l'analyse dès l'étape 3.

3.12.2 REPRÉSENTATION GRAPHIQUE DES RÉSIDUS

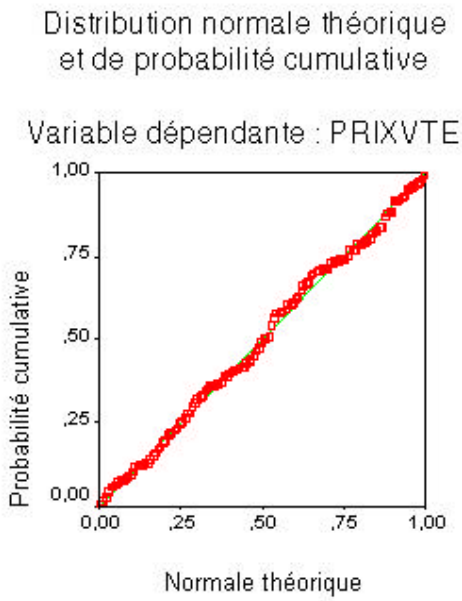
L'analyse des résidus doit toujours s'accompagner des graphiques appropriés : Il s'agit notamment du graphique des résidus standardisés, dans lequel la courbe normale se superpose aux histogrammes, et celui permettant de confronter la distribution de probabilité cumulative des résidus standardisés avec la distribution normale théorique.

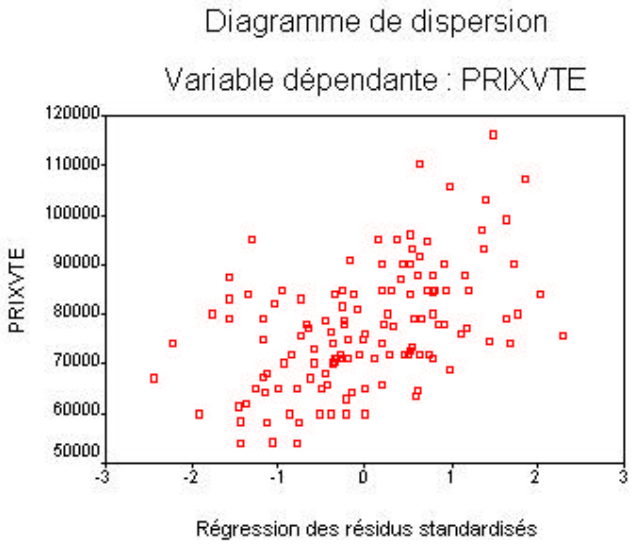
Tableau 3.11 - Présentation graphique des résidus

Dans le premier graphique, la courbe normale se superpose à l'histogramme. Ce dernier reproduit les indicateurs statistiques de la seconde partie du tableau 3.10 sur la structure des résidus standardisés, dont l'écart type se situe ici à 0,96 en dépit de la présence de résidus extrêmes. Il importe de rappeler qu'en vertu d'une distribution parfaitement normale, les résidus standardisés de la régression sont de moyenne *« 0 »* et d'écart type *« 1 »* (illustré par le second graphique). En conclusion, dans le cas présent, la structure des termes d'erreur se rapproche de la distribution normale et, par conséquent, ne pose pas de problèmes.



Le deuxième graphique représente la distribution de probabilité cumulative des résidus standardisés (voir les carreaux) du modèle, en fonction de la distribution normale théorique des prix de vente (droite diagonale pleine). Ce graphique sert, notamment, à juger de l'ampleur du phénomène d'hétéroscédasticité des erreurs, fréquent dans les modèles d'évaluation (voir le point 3.4.4.6 de ce chapitre). Ce problème, faut-il le rappeler, se produit lorsque la *variance* des résidus, donc leur dispersion autour de la moyenne, n'est pas constante sur l'ensemble de l'échantillon et augmente avec la valeur de la variable dépendante. Cela se traduit graphiquement par un nuage de points, en forme de « trompette de Jéricho », c.-à-d. évasé vers la droite. Bien que ce ne soit visiblement pas le cas ici, il est possible de constater que les résidus standardisés du modèle statistique se distribuent selon un axe dont la pente est légèrement positive. Cela reflète une réalité assez systématique en évaluation immobilière, à savoir que les propriétés « bas de gamme » (60 000 $ et moins) sont en général surestimées par l'équation de régression (résidus négatifs), alors que les résidences « haut de gamme » (100 000 $ et plus) sont, au contraire, sous-estimées (résidus positifs).





En conclusion de cette étape, l'analyse des résidus constitue une étape essentielle en modélisation statistique, dans la mesure où elle fournit une information riche et diversifiée sur la nature de l'échantillon utilisé et sur les failles empiriques (lacunes dans la construction de la base de données) et méthodologiques (omission de variables clés, mauvaise spécification de la forme fonctionnelle de l'équation) qui auraient pu échapper à l'analyste lors des étapes antérieures. Toute lacune majeure se répercute sur la structure des résidus, affectant ainsi les performances du modèle et la fiabilité des prix modélisés. C'est donc l'étape permettant de « corriger le tir » et d'arriver à la mise au point du modèle définitif qui fait l'objet de l'*étape 8* présentée aux pages suivantes.

3.13 ÉTAPE 8 : MISE AU POINT DU MODÈLE FINAL

À la suite de l'analyse des résidus de la régression, cette étape permet de procéder à la mise au point de la version finale du modèle statistique. En règle générale, et dans la mesure où l'étape antérieure n'a pas mis à jour de problèmes majeurs impliquant une recalibration en profondeur de l'équation de régression, la spécification du modèle obtenu à l'étape 6, par le biais de la procédure *Stepwise,* sera reprise intégralement à l'aide de la procédure *Enter*. Évidemment, les observations générant les résidus extrêmes identifiés à l'étape 7 seront, au préalable, retirés de l'analyse. Pour y parvenir, il sera approprié de recourir à la même procédure que celle ayant servi à la sélection d'un sous-échantillon pour fins de validation ultérieure du modèle (voir le point 3.9.2 de ce chapitre).

3.13.1 EXEMPLE D'APPLICATION

Les résultats de cette opération figurent au *tableau 3.12* présenté à la page suivante. Le modèle final comporte donc onze variables explicatives et 128 observations.

a) Validation du modèle final

À la suite du retrait des ventes n^{os} *119*, *135*, *146* et *151*, le nouveau prix moyen de l'échantillon épuré s'établit à 77 318 $ (voir les statistiques descriptives en fin de tableau). Comme il est possible de le constater, le « R^2 » ajusté est passé à près de *78 %,* une amélioration substantielle par rapport aux résultats du tableau 3.9 (73,2 %), alors que l'erreur type de prévision a chuté à 5 703 $ (versus 6 198 $ antérieurement); l'erreur d'estimation relative (CV) s'établit donc à *7,38 %* du prix moyen, ce qui est excellent. Le test *« F »* enfin, affichant une valeur de *41,46*, en ressort également grandement renforcé.

b) Interprétation des prix modélisés

L'analyse des coefficients de régression, à la suite du retrait des résidus extrêmes, vient confirmer la très grande stabilité du modèle statistique. Les onze paramètres sont tout à fait conformes aux attentes, en amplitude comme en signe, et sont, par ailleurs, tous significatifs au seuil de 0,01 (à l'exception du coefficient de la variable AIRFINSS, qui l'est au seuil 0,02). Quant à la multicolinéarité, elle est, à toutes fins pratiques, absente de l'équation, le VIF le plus élevé demeurant inférieur à 1,5. Sur cette base, il devient donc possible d'interpréter les prix modélisés en toute confiance. Ainsi, il est important de retenir :

- que chaque année d'âge apparent réduit la valeur d'une unité d'évaluation d'un peu plus de 700 $ (soit une dépréciation annuelle de 0,9 %);

- que chaque m² d'aire habitable aux étages apporte une contribution additionnelle de 353 $ à la valeur marchande, alors que cette contribution n'est que de 43 $, soit près de sept fois moins, dans le cas des espaces au sous-sol;

- qu'une unité d'évaluation de classe 4 obtienne une prime de marché de quelque 5 400 $ par rapport à celles de classe 5;

- que la présence d'une salle de bain supplémentaire représente une valeur additionnelle de 3 400 $;

- qu'une cuisine moderne et fonctionnelle ajoute quelque 2 000 $ à la valeur, alors que la présence d'une thermopompe apporte une contribution d'un peu plus de 6 700 $;

- que chaque espace de garage est évalué par le marché à 3 300 $;

- qu'une piscine excavée a une valeur marchande de 9 500 $, contre 3 400 $ environ pour une piscine hors terre;

- enfin, que chaque m² de terrain apporte une contribution marginale additionnelle de 4,30 $.

Il importe de rappeler ici que ces prix modélisés ne s'appliquent, en principe, qu'à l'univers des unités d'évaluation visées par l'échantillon servant à la calibration du modèle; en l'occurrence, il s'agit des « bungalows » situés sur le territoire de la ville sous étude. Il est déconseillé de les utiliser dans un autre contexte et cela ne doit se faire que moyennant une mise en garde explicite sur les risques encourus.

Tableau 3.12 - Modèle final - Régression standard (mode Enter)
K (var. indép.) = 11
N = 128 observations

Modèle	Variables utilisées	Variables retirées	Méthode
1	SUPTER, PISCHT, CLASSUP, THERMOPO, SDBAIN, CUISEXTR, NBPLGAR, PISCEXC, AIREHAB, AIRFINSS, AGEAPP[a]	,	Enter

a. Ensemble des variables utilisées pour la requête.

Sommaire du modèle

Modèle	R	R carré	R carré ajusté	Erreur type
1	,893[a]	,797	,778	$5,703

a. Predictors: (Constant), SUPTER, PISCHT, CLASSUP, THERMOPO, SDBAIN, CUISEXTR, NBPLGAR, PISCEXC, AIREHAB, AIRFINSS, AGEAPP

Analyse de variance[b]

Modèle		Somme des carrés	df	Moyenne des carrés	F	Sig.
1	Régression	1,483E+10	11	1348509835,9	41,459	,000[a]
	Residu	3773019754	116	32526032,362		
	Total	1,861E+10	127			

a. Prévision : (Constante), SUPTER, PISCHT, CLASSUP, THERMOPO, SDBAIN, CUISEXTR, NBPLGAR, PISCEXC, AIREHAB, AIRFINSS, AGEAPP

b. Variable dépendante : PRIXVTE

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés			Statistiques de colinéarité	
		B	Erreur type	Beta	t	Sig.	Tolérance	VIF
1	(Constante)	$38,709	$4,789		8.083	.000		
	AGEAPP	-$717	$79	-.445	-9.055	.000	.724	1.381
	AIREHAB	$353	$47	.365	7.477	.000	.733	1.364
	AIRFINSS	$43	$18	.116	2.392	.018	.748	1.336
	CLASSUP	$5,398	$1,517	.178	3.559	.001	.703	1.423
	SDBAIN	$3,422	$1,178	.141	2.903	.004	.738	1.354
	CUISEXTR	$2,015	$673	.138	2.994	.003	.819	1.221
	THERMOPO	$6,708	$2,205	.135	3.043	.003	.892	1.121
	NBPLGAR	$3,308	$818	.192	4.042	.000	.775	1.290
	PISCEXC	$9,523	$2,194	.212	4.340	.000	.733	1.365
	PISCHT	$3,375	$1,184	.126	2.850	.005	.896	1.116
	SUPTER	$4.28	$1.24	.158	3.439	.001	.825	1.212

a. Variable dépendante : PRIXVTE

Statistiques descriptives

	N	Minimum	Maximum	Moyenne	Écart type
PRIXVTE	128	$54,000	$116,000	$77,318	$12,104
Valid N (listwise)	128				

En dépit des performances plus qu'honnêtes que produit la version finale du modèle de régression, une dernière validation s'impose. Elle consiste à soumettre le modèle à un test ultime de robustesse, à savoir *sa capacité d'estimer correctement la valeur des unités d'évaluation totalement exogènes à l'exercice de modélisation*. C'est précisément l'objet de l'*étape 9* présentée aux pages suivantes.

3.14 ÉTAPE 9 : VALIDATION DU MODÈLE FINAL

La validation du modèle final consiste à appliquer, aux observations du sous-échantillon (voir le point 3.9.2 de ce chapitre), les paramètres de l'équation de régression. Pour chaque immeuble du sous-échantillon, il convient alors de multiplier la valeur attribuée à chacune des variables explicatives retenues par le coefficient de régression correspondant. La valeur estimée par le modèle est obtenue en effectuant la somme des produits ainsi obtenus, à laquelle s'ajoute, évidemment, le terme constant (l'ordonnée à l'origine). La valeur estimée est ensuite déduite du prix de vente observé et le résidu en découlant se traduit en pourcentage du prix, ce qui permet d'établir *l'erreur d'estimation relative,* exprimée en valeur absolue. Il suffit, enfin, de calculer la moyenne arithmétique de ces erreurs d'estimation relatives, pour obtenir la performance prédictive du modèle statistique lorsqu'il est appliqué à des données indépendantes (mais tirées de l'échantillon initial, donc représentatif de la population étudiée). La robustesse du modèle est assurée si la performance *du sous-échantillon* est conforme à la performance *de l'échantillon*, telle qu'établie par l'erreur type relative (CV).

3.14.1 EXEMPLE D'APPLICATION

Le *tableau 3.12* suivant reproduit cet exercice à l'aide des quatorze ventes du sous-échantillon retenu. Comme il est possible de le constater, l'erreur d'estimation est de 6,3 % ou moins pour douze de ces immeubles; dix transactions produisent, par ailleurs, une erreur de 4,7 % ou moins, alors que dans six cas, l'erreur de prévision est inférieure à 3,0 %. Quant à l'erreur moyenne d'estimation du modèle, elle est de 5,0 %, une excellente performance ne laissant aucun doute sur la robustesse du modèle de régression. Ceci dit, deux ventes génèrent une erreur de plus de 10,0 %. Dans le cas de la vente *n° 128* (sous-estimée par le modèle), elle s'établit à 13,0 %, ce qui n'est pas excessif. Dans le cas de la vente *n° 32* (surestimée par le modèle), l'erreur atteint 12,5 %. Une analyse plus approfondie de ces cas marginaux est donc appropriée (voir le point 3.14.2 de ce chapitre).

Constante:	*38709*											
Coefficient de régression:			*-717*	*353*	*43*	*5398*	*3422*	*2015*	*6708*	*3308*	*9523*	*3375*	*4,32*
	NOVTE	**PRIXVTE**	**AGEAPP**	**AIREHAB**	**AIRFINSS**	**CLASSUP**	**SDBAIN**	**CUISEXTR**	**THERMOPO**	**NBPLGAR**	**PISCEXC**	**PISCHT**	**SUPTER**
1	10	$65 000	28	96	87	0	1	0	0	0	0	0	836
2	20	$65 000	29	112	56	0	1	0	0	0	0	1	627
3	32	$68 500	15	94	94	0	2	0	0	0	0	1	678
4	43	$68 000	29	102	26	0	1	0	0	1	0	0	836
5	54	$72 500	15	93	84	0	1	0	0	0	0	0	697
6	64	$84 500	27	125	111	0	2	0	0	1	0	0	929
7	74	$71 500	18	91	43	0	1	0	0	0	0	1	760
8	84	$65 000	23	109	0	0	1	0	0	0	0	0	837
9	95	$69 000	21	94	0	0	1	0	0	0	0	1	790
10	107	$79 000	4	89	33	1	2	0	0	0	0	0	451
11	118	$62 000	29	107	64	0	1	0	0	0	0	0	664
12	128	$78 000	25	97	0	0	1	0	0	2	0	0	917
13	138	$78 000	12	104	13	0	1	0	0	0	0	1	1394
14	148	$92 500	6	105	0	0	1	0	0	1	0	1	1394

Prix estimé	**Résidu ($)**	**Résidu (%)**
63 296	1 704	**2,7%**
69 366	-4 366	**6,3%**
78 326	-9 826	**12,5%**
65 382	2 618	**4,0%**
70 828	1 672	**2,4%**
82 413	2 087	**2,5%**
69 855	1 645	**2,4%**
67 733	-2 733	**4,0%**
67 044	1 956	**2,9%**
82 867	-3 867	**4,7%**
64 729	-2 729	**4,2%**
69 024	8 976	**13,0%**
80 195	-2 195	**2,7%**
87 599	4 901	**5,6%**

Erreur moy. de prévision:	***Tous les cas:***	**5,0%**
	Sans vte 32 et après ajust.:	***3,6%***

3.14.2 ANALYSE DES CAS MARGINAUX

L'analyse des cas marginaux permet de constater que les erreurs d'estimation du modèle sont dues soit à des informations manquantes qu'il est possible de réintégrer dans l'analyse (vente *n° 128*, ajout d'un garage), soit à des éléments relatifs à la propriété (avant-toit de la galerie avant dans le cas de la vente *n° 32*) ou au secteur : zone enclavée par le parc industriel (vente *n° 32*); présence dans le voisinage immédiat de propriétés luxueuses (vente *n° 128*). En retranchant du calcul la vente *n° 32* et en apportant à la vente *n° 128* l'ajustement approprié (soit 6 616 $), l'erreur moyenne d'estimation du modèle chute à *3,6 %*.

N° de vente	Adresse civique	Prix de vente/Évaluation/ Valeur estimée	Commentaires
32	845, Des Grives	Prix de vente : 68 500 $ Évaluation : 74 800 $ Val. estimée : 78 109 $	• L'immeuble est situé dans un secteur de statut socio-économique inférieur et enclavé par le parc industriel; en outre, l'avant-toit de la galerie avant est très avancé, ce qui assombrit l'intérieur de la propriété. • Le modèle ne tient compte d'aucun de ces éléments.
128	30, rue des Pins	Prix de vente : 78 000 $ Évaluation : 77 600 $ Val. estimée : 68 731 $	• Outre le fait que l'immeuble a fait l'objet d'un ajout (un garage) n'ayant pas été intégré au rôle d'évaluation, celui-ci fait face à trois propriétés luxueuses rehaussant la valeur des unités d'évaluation avoisinantes. • En tenant compte de l'information manquante relative au garage, l'erreur d'estimation passe donc de 9 269 $ à 2 653 $, soit 3,4 % du prix de vente.

3.14.3 CONCLUSION

Aucun modèle statistique, si performant soit-il, ne peut prédire avec une précision extrême *toutes* les valeurs marchandes des unités d'évaluation d'un segment de marché donné. Cela n'est, d'ailleurs, pas son but, puisque la modélisation vise *à généraliser le processus de formation des prix résidentiels* à partir des attributs les plus déterminants. Il permet plutôt d'estimer ces valeurs de façon optimale, en considérant l'ensemble du marché local plutôt que des cas isolés et de le faire *de façon beaucoup plus efficiente* que ne le font les méthodes traditionnelles, puisque le nombre de descripteurs utilisés se trouve réduit au strict minimum. Il permet, d'autre part, d'atteindre ce qui constitue l'objectif premier d'un rôle d'évaluation de qualité, à savoir *son caractère équitable*, dans la mesure où toutes les unités d'évaluation d'un même secteur d'analyse sont évaluées de façon uniforme, c.-à-d. sur la base des mêmes critères. Quant aux cas marginaux, étant par définition très restreints en nombre, ils peuvent faire l'objet d'une attention particulière et être traités séparément. Les rajustements à apporter proviendront alors soit d'analyses parallèles menées à l'aide de méthodes traditionnelles (comparaisons par pair, jugement de l'analyste), soit, préférablement, de résultats obtenus à l'aide de modèles statistiques portant sur un autre échantillon jugé relativement comparable. Bien que la prudence soit de mise ici, il est possible de constater, en effet, que la contribution marginale *relative* de plusieurs attributs résidentiels demeure assez constante d'un marché à l'autre[1]. Cette approche, bien que ne donnant qu'une approximation de l'impact recherché, demeure néanmoins préférable à une évaluation totalement subjective ou arbitraire du phénomène.

[1] À titre d'exemple, s'il était possible de disposer, pour une autre ville québécoise de taille et de vocation similaire de celle sous étude, d'un modèle permettant d'isoler la baisse de valeur attribuée à l'absence de services publics, cela permettrait, sans grand risque, d'appliquer le coefficient obtenu, exprimé en pourcentage du prix de vente moyen, au cas étudié.

3.15 ÉTAPE 10 : PRODUCTION D'INDICATIONS DE LA VALEUR

Lorsque la technique de modélisation statistique est appliquée à des fins d'évaluation foncière municipale, les unités d'évaluation d'une même unité de voisinage sont soumises aux mêmes critères d'évaluation que l'ensemble des autres unités d'évaluation. Toutefois, le modèle statistique ne doit s'appliquer qu'en fonction de son domaine de validation et de ses limites.

Aussi, tant pour l'évaluateur municipal que pour son client (la municipalité), cette mise au point est importante, car elle réduit les risques d'une utilisation inadéquate des résultats de l'analyse et les erreurs susceptibles d'en affecter l'interprétation. Il sera approprié, en l'occurrence, de rappeler à l'utilisateur que le modèle n'atteint ses performances optimales que dans la mesure où il est appliqué à l'univers auquel il est prioritairement destiné. À cet égard, en étendre les conclusions à des unités d'évaluation appartenant à un parc résidentiel, dont la structure et les caractéristiques diffèrent sensiblement de celles de l'échantillon utilisé pour le construire, ne peut qu'entraîner des estimations erronées. Il est également prudent de ne pas interpréter les prix modélisés obtenus en dehors de leur contexte de modélisation, c.-à-d. sans tenir compte des influences croisées entre les variables explicatives, influences sur lesquelles ils sont fondés et pouvant différer d'un marché à l'autre.

Enfin, l'analyste doit faire mention de toute limite ou lacune connue, en ce qui a trait à la nature de l'information utilisée pour modéliser son marché et nuancer, en conséquence, son interprétation de la performance globale du modèle et de la validité des coefficients de régression, établissant la contribution marginale de chaque attribut de l'unité d'évaluation à sa valeur marchande.

3.15.1 DOMAINE DE VALIDATION ET LIMITES

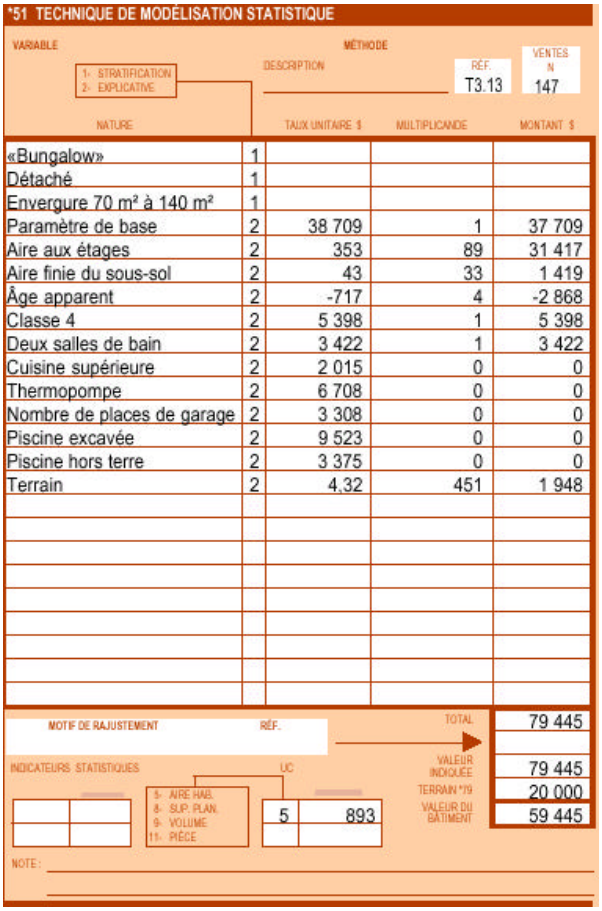
Les paramètres élaborés pour le modèle ne doivent s'appliquer qu'aux « bungalows » situés sur le territoire de la ville sous étude.

Les paramètres ne s'appliquent qu'aux bâtiments de classe 4 ou 5, de type détaché, dont la date de construction apparente n'excède pas 30 ans et que l'aire habitable se situe entre 70 et 140 m².

3.15.2 EXEMPLE D'APPLICATION

Le bloc *51 de la fiche de propriété 2.6.1 C permet de consigner les résultats de l'application de la technique de modélisation statistique pour une unité d'évaluation sous évaluation. L'exemple de la page suivante montre une application de ce bloc. La partie II de ce volume donne les explications concernant l'utilisation de cette fiche de propriété.

Tableau 3.13 TECHNIQUE DE MODÉLISATION STATISTIQUE				
Description de l'immeuble			Paramètre	Valeur
1. Dimension du bâtiment (base)			38709	38709
Aire aux étages (m²)		89	353	31417
Aire finie au sous-sol (m²)		33	43	1419
2. Date de construction				
Âge apparent		4	-717	-2868
3. Classe		4	5398	5398
4. Nombre de salles de bain		2	3422	3422
5. Cuisine supérieure		non	2015	0
6. Thermopompe		non	6708	0
7. Nombre de places de garage		0	3308	0
8. Piscine				
Excavée		non	9523	0
Hors terre		non	3375	0
9. Dimension du terrain (m²)		451	4,32	1948
Valeur totale				**79445**



*51 TECHNIQUE DE MODÉLISATION STATISTIQUE

VARIABLE — 1- STRATIFICATION 2- EXPLICATIVE — MÉTHODE — DESCRIPTION — RÉF. T3.13 — VENTES N 147

NATURE		TAUX UNITAIRE $	MULTIPLICANDE	MONTANT $
«Bungalow»	1			
Détaché	1			
Envergure 70 m² à 140 m²	1			
Paramètre de base	2	38 709	1	37 709
Aire aux étages	2	353	89	31 417
Aire finie du sous-sol	2	43	33	1 419
Âge apparent	2	-717	4	-2 868
Classe 4	2	5 398	1	5 398
Deux salles de bain	2	3 422	1	3 422
Cuisine supérieure	2	2 015	0	0
Thermopompe	2	6 708	0	0
Nombre de places de garage	2	3 308	0	0
Piscine excavée	2	9 523	0	0
Piscine hors terre	2	3 375	0	0
Terrain	2	4,32	451	1 948

MOTIF DE RAJUSTEMENT — RÉF. — TOTAL 79 445

INDICATEURS STATISTIQUES — 5- AIRE HAB. 8- SUP. PLAN. 9- VOLUME 11- PIÈCE — UC 5 893

VALEUR INDIQUÉE 79 445 — TERRAIN *79 20 000 — VALEUR DU BÂTIMENT 59 445

NOTE :