Évaluation de la valeur foncière par l'emploi de la DVF une application au cas de la ville de Caen

Amaury Artault, Benjamin Giot et Arthur Gires

Université de Caen Normandie, Campus II, Science 3, 14032, Caen, France

14 mars 2021

Abstract

L'objectif de ce papier est d'utiliser les données "Demande de Valeurs Foncières" pour expliquer la valeur foncière des logements sur la commune de Caen. Ces données sont complétées par diverses sources afin de gagner en précision. Une régression permet de mettre en évidence de multiples variables explicatives. Le modèle permet alors d'avoir une première idée de la valeur foncière d'un logement et des facteurs majeurs l'expliquant, tout en ayant une complexité raisonnable. Une analyse en composantes principales complétée par une classification dégage une certaine homogénéité de la valeur foncière sur l'ensemble de la ville normande.

1 Introduction

Chaque mois, ce sont des centaines de ventes immobilières qui se produisent sur la commune de Caen. Souvent, cela reste un évènement inaccoutumé que ce soit pour l'acquéreur ou pour le vendeur. Ainsi, il n'est pas rare que ces deux acteurs émettent une incertitude quant à l'estimation du bien. En effet, les facteurs rentrant en jeu dans l'évaluation de la valeur foncière peuvent être nombreux. Certains sont criants tels que la surface du logement mais d'autres restent relativement plus discrets. Il est alors difficile de tout prendre en compte et de juger de l'importance de chaque critère.

Depuis mai 2019, les données "Demande de Valeurs Foncières" (DVF) sont disponibles en open-data sur le site data.gouv.fr. Il y est indiqué que ce jeu de données est publié et produit par la direction générale des finances publiques. Il permet de connaître les transactions immobilières intervenues au cours des cinq dernières années sur le territoire métropolitain et les DOM-TOM, à l'exception de l'Alsace, de la Moselle et de Mayotte. Les données contenues sont issues des actes notariés et des informations cadastrales.

Ainsi, il est intéressant et possible de se questionner sur le potentiel des données DVF afin d'estimer de manière fiable la valeur foncière d'un bien immobilier. L'objectif de ce papier est donc d'exploiter cette base de données afin de répondre à cette interrogation en déterminant des variables explicatives et en décelant des profils de logements sur la commune de Caen. Par conséquent, après avoir exposé la construction de la base utilisée et son contenu final, une régression est appliquée afin de juger de la pertinence de chaque variable. Puis, une analyse en composantes principales suivi d'une classification est réalisée dans le but de dégager des groupes dépendant des variables.

Cette étude tire son inspiration de deux articles d'économie et statistique. Le premier, intitulé "Évaluation des aménités urbaines par la méthode des prix hédoniques : une application au cas de la ville d'Angers" et produit par Muriel Travers, Gildas Appéré et Solène Larue, traite l'implication des aménités urbaines dans le choix résidentiels des ménages. Le second nommé "Le prix des attributs du logement" et rédigé par Jean Cavailhès s'intéresse aux attributs du logement pour expliquer le prix global au moyen de la méthode des prix hédonistes. Ce papier s'engage alors sur un sujet similaire à l'aide des outils statistiques vus lors de la formation proposée par le master SAAD à Caen.

2 Élaboration de la base de données

2.1 Les données DVF

Depuis 2019, les données "Demande de Valeurs Foncières" (DVF) sont publiées par la direction générale des Finances publiques. Toutes les transactions foncières et immobilières effectuées en France durant ces cinq dernières années sont alors référencées dans un fichier accessible librement. Chaque mutation foncière ou immobilère est sujet à un enregistrement de la part de notaires. Pour des raisons fiscales et légales, les documents concernant ces transferts d'un bien sont transmis à l'administration nationale qui alimente la base de données nommée "Base nationale des données patrimoniales" (BNDP). Les données DVF sont issues d'une extraction de la base BNDP. Cependant, les informations sur les mutations y sont enrichies par des informations cadastrales. Ainsi, pour chaque transaction, des informations économiques telle que la valeur, des informations cadastrales comme le numéro de parcelle et des informations sur le bien avec par exemple sa surface. Toutes les mutations sont anonymisées, datées et localisées au moyen de l'adresse. À ce jour, les données DVF établissent la liste la plus exhaustive dans le recensement des mutations foncières et immobilières en France. Chaque année, ce sont près de trois millions de transactions qui sont publiques.

Néanmoins, les données DVF recensent seulement les transferts de biens onéreux tels que les ventes, les échanges, les expropriations et les adjudications. Les informations sur les successions et les donations n'y figurent pas. De plus, les données peuvent être incomplètes sur les périodes les plus récentes. Une actualisation des données DVF est réalisée tous les six mois pour pallier cette contrainte. Une autre importante difficulté est due à la présence de multilignes. En effet, une même mutation peut être renseignée sur plusieurs lignes. Cela est le cas lorsque la transaction concerne plusieurs biens ou lorsqu'un bien est localisé à travers plusieurs parcelles cadastrales. Le cas de multilignes concernerait près de la moitié des mutations. Afin d'en réaliser une analyse, il est donc important de dédoublonner les données.

2.2 La base DVF+

Une solution possible à la contrainte des multilignes est l'utilisation de la base de données "DVF+ open-data" proposé par le Cerema. Le Cerema est le centre d'études et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement. C'est un établissement public tourné vers l'appui aux politiques publiques et dépendant du Ministère de la transition écologique et du ministère de la cohésion des territoires et des relations avec les collectivités territoriales. La DVF+ permet d'accéder aux données DVF sous un modèle structuré, géolocalisé et rendant plus simple l'exploitation des diverses données. Dans cette base, chaque mutation est résumée dans une seule ligne. Par ailleurs, alors que les données DVF sont construites avant tout pour des raisons fiscales, la DVF+ ajoute des variables issues des données brutes mais transformées ou agrégées pour permettre une analyse plus aisée du caractère foncier et immobilier des transactions.

Dans cet étude, la DVF+ est donc utilisée comme base principale préférablement aux données DVF car elle permet une exploitation simplifiée des données. Dénombrant 58 variables, toutes ne sont pas utiles pour l'analyse recherchée dans ce papier. Tout d'abord, seul le fichier consacré aux mutations dans le Calvados sera considéré. Cela permet de s'exempter du coté volumineux caractéristique aux données DVF à portée nationale et de garder seulement les transactions spécifiques à la ville de Caen. Les variables à propos de la date de la mutation, de la valeur, du type de bien et de la surface bâtie et de la surface du terrain sont prises en compte. En revanche, les variables spécifiques et exclusives à chaque type de bien sont exclues de part une volonté d'analyse générale de la valeur foncière. Cela correspond par exemple aux nombres d'appartment à 4 pièces principales ou bien la surface de l'ensemble des maisons avec 3 pièces principales. Toutes les ventes en l'état futur d'achèvement sont retranchées de l'étude car la DVF+ n'y associe pas de surface définie. Enfin, les informations cadastrales et de géolocalisations ne sont pas retenues non plus. En effet, ni le plan cadastral à finalité fiscale ni la géolocalisation trop précise n'est efficace dans la détermination de la valeur foncière dans le cas présent. Idéalement, cette donnée précise aurait pu être utilisée. Cependant afin de simplifier l'analyse, le découpage en IRIS est préféré et offre plus d'opportunité d'analyse par une disponibilité de données externes par IRIS.

2.3 Découpage en IRIS

Développé par l'institut national de la statistique et des études économiques (Insee) dès 1999 pour le recensement de la population, l'IRIS, signifiant "Ilots Regroupés pour l'Information Statistique", est dorénavant la référence en terme de diffusion de données infra-communales. Il est indiqué sur le site de l'Insee que l'IRIS correspond à un découpage en mailles de tailles homogènes du territoire français répondant à des critères géographiques et démographiques tout en étant stable dans le temps et facilement identifiable cartographiquement. Il existe trois type d'IRIS :

- les IRIS d'habitat représentant la majorité des IRIS et caractérisés par une population se situant entre 1800 et 5000 habitants et une homogénéité quant au type d'habitat ;
- les IRIS d'activité distingués par un regroupement de plus de 1000 salariés et un nombre deux fois plus élevé d'emploi salariés que de population résidente ;
- les IRIS divers particularisés par de grande superficie peu habitée.

La ville de Caen est composée d'un total de 51 IRIS, comprenant 46 IRIS d'habitat, 3 IRIS d'activité et enfin 2 IRIS divers. Le découpage en IRIS a un double intérêt dans cette étude. Tout d'abord, il permet un découpage plus homogène de la ville de Caen que le plan cadastral et habilite l'observation de tendances possibles selon des zones pertinentes au sein de la commune. Ensuite, comme exposé précédemment, l'Insee utilise le découpage en Iris dans de nombreuses études et cela permet donc d'agrémenter la base DVF+ de diverses variables supplémentaires hypothétiquement judicieuses à l'étude de la valeur foncière.

L'obtention de l'IRIS de chaque mutation se fait au moyen du site "Pyris : Insee IRIS Geolocalizer". Celui-ci permet d'avoir l'IRIS en partant d'une adresse postale. Il est basé sur l'utilisation d'un système de géolocalisation par coordonnées issues de l'interface de programmation (API) de la Base Adresse Nationa (BAN). La BAN est une base de données nationale du service public des données de référence. L'adresse postale précise de chaque transaction est récupérable dans les données DVF. Ainsi, grâce à un programme réalisé avec le logiciel R, une requête de l'API est effectuée pour chaque mutation localisée sur la ville de Caen. Il en résulte des informations à propos de l'IRIS telles que le code, le nom et le type, qui sont reliées à la date et la valeur de la mutation dans une base de données concise. Une jointure est ensuite réalisée avec le logiciel Talend entre cette nouvelle base de données et celle principalement utilisée dans l'étude à savoir la DVF+ sur les variables date et valeur de la transaction.

Maintenant et comme précédemment évoqué, il devient plus aisé d'enrichir la base DVF+ avec des données établies sur la commune de Caen, distinguées par IRIS et trouvables sur le site de l'Insee.

2.4 Ajout de données découpées en IRIS

Une donnée qui peut être intéressante pour l'évaluation de la valeur foncière d'un bien est la présence et le nombre d'équipements tels que des commerces, des services de santé ou des écoles à proximité. Il est envisageable que la présence proche de nombreux équipements a un impact positif sur la valeur foncière. Dans le cas de l'étude, la proximité se traduit par la présence dans le même IRIS. Disponible sur le site de l'Insee, la base permanente des équipements (BPE) dénombre les équipements par IRIS et par type réunis sous 7 thèmes principaux et distincts :

- service aux particuliers (public, général, ...) ;
- commerce (alimentaire et non alimentaires) ;
- enseignement ;
- santé et social ;
- transports et déplacements ;
- sport loisirs et culture ;
- tourisme.

Dans le cadre de l'analyse présentée dans ce papier, une restructuration de ces catégories est effectuée afin de réduire le nombre de variables étudiées et par conséquent la complexité de l'étude. Les thèmes "commerce", "service aux particuliers" et "transports et déplacements" sont de la sorte rattachés . Le thème" sport, loisirs et culture" est ainsi réuni avec le thème "tourisme" sous une seule variable. Enfin, le thème "enseignement" est relié au thème "santé

et social".

Le revenu médian déclaré par unité de consommation est de même une donnée pertinente, fournie par l'Insee et cataloguée par IRIS. Il est probable que la valeur foncière d'un bien situé dans un IRIS est d'autant plus élevé que le niveau de vie de la population habitant l'IRIS. Cette donnée est manquante pour certains IRIS et est donc substituée par le revenu médian déclaré par unité de consommation de la ville de Caen.

Partageant plus ou moins la même hypothèse que la donnée précédente, il peut se relever judicieux de prendre en compte les proportions relatives aux catégories socioprofessionnelles, disponibles sur le site de l'Insee et référencées par IRIS. Seulement la proportion de personnes "cadres et de profesions intellectuelles supérieures" parmi la population active est ici ajoutée à la base principale par soucis de simplicité et de redondance. Cela permet d'avoir un autre indicateur basé sur le niveau de vie.

2.5 Ajout de variables supplémentaires

La base de données utilisée dans l'étude basée sur la DVF+ comporte dorénavant des informations par IRIS venant de bases externes. Cependant, il est de même possible d'ajouter des variables créées à partir de celles déjà présentes afin d'établir des données plus pertinente ou plus maniables. Une première variable qualitative nommée jardin est ainsi réalisée. Elle traduit la présence ou non d'un jardin s'appuyant sur le surface du terrain. Une variable qualitative année est aussi constituée afin de déceler une éventuelle évolution annuelle de la valeur foncière. La variable date est délaissée car bien trop précise et non pertinente.

Enfin, le nombre d'IRIS sur la ville de Caen étant important, une classification des IRIS selon leur localisation est effectuée afin de simplifier ce facteur de l'étude. Quatre catégories sont établis pour cette variable qualitative zone_iris : "Centre", "Bordure", "Bordure-entre" et "Entre-deux". La modalité de chaque IRIS est identifiée à partir de la cartographie des IRIS disponible sur le site https ://www.geoportail.gouv.fr. Cela peut permettre de donner une réponse quant à l'interaction possible entre la valeur foncière et la proximité du centre-villle à Caen. Près de 25% des mutations sont identifiées en "Bordure", 20% en "Centre", 37% en "Entre-deux" et 17% en "Bordure-entre".

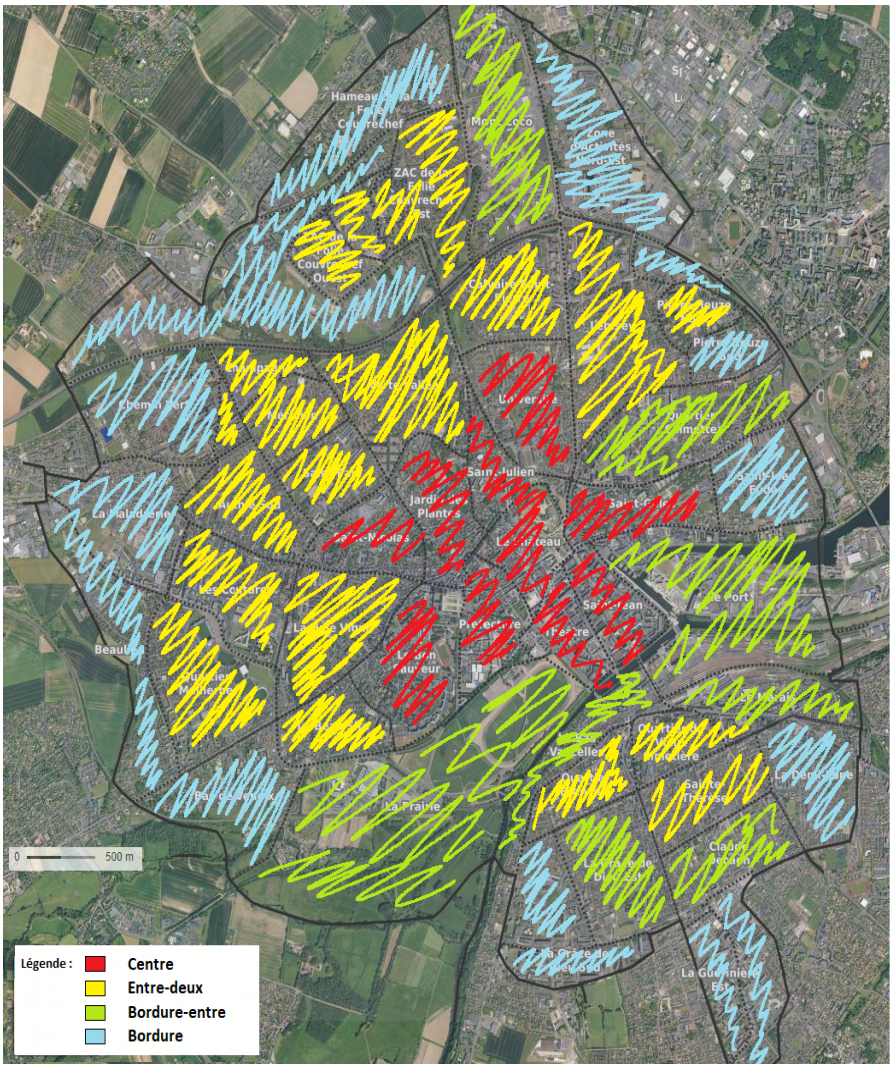

FIGURE 1 : Carte de la répartition en zone des IRIS.

Pour finir, les données sont réduites sur les mutations concernant un appartement ou une maison, excluant donc les transactions de terrains, d'activités et de maisons ou appartements multiples qui de part des caractéristiques plus distinctes peuvent être amenées à fausser les résultats d'une analyse qui se veut donc portée sur la population habitante. Ensuite, une élimination de diverses valeurs anormales est entreprise. Les mutations dont la valeur ne dépasse pas les 10 000 euros et celles dont la surface du bâtiment est inférieur à 10 mètres carrés sont écartées de la base de données. Il en est de même pour les transactions avec un rapport entre la valeur foncière et la surface du bâtiment n'atteignant pas 500, ce qui correspond de manière grossière à un prix de 500 euros par mètre carré.

Ainsi, la base de données finale obtenue comporte 6429 mutations de maison ou d'appartement, chacune défini par 15 variables, s'étalant sur un période allant de 2016 à 2020 et localisé sur la ville de Caen. Néanmoins, il est à considérer que pour les années 2019 et 2020, le nombre de transactions est bien plus faible que les années qui les précédent. Ce caractère incomplet est lié à la fraîcheur des données pour lesquelles l'actualisation semestrielle est

toujours nécessaire.

2.6 Dictionnaire des variables

N°	Nom	Description	Type
1	valeur	Valeur foncière du bien de la mutation	Quantitative
2	sbati	Surface de l'ensemble du bâti ayant muté	Quantitative
3	sterr	Surface de terrain ayant muté	Quantitative
4	jardin	Présence ou non d'un jardin	Qualitative
5	typebien	Type du bien : maison ou appartement	Qualitative
6	annee	Année effective de la mutation	Qualitative
7	code_iris	Code Insee de l'IRIS	Qualitative
8	nom_iris	Nom de l'IRIS	Qualitative
9	type_iris	Type de l'IRIS	Qualitative
10	zone_iris	Zone de l'IRIS selon sa localisation	Qualitative
11	rmedian	Revenu déclaré médian par unité de consommation	Quantitative
12	csp	Proportion de cadres et professions intellectuelles supérieures	Quantitative
13	serv_equip	Nombre d'équipements service aux particuliers et commerce	Quantitative
14	soc_equip	Nombre d'équipements enseignement, santé et social	Quantitative
15	slc_equip	Nombre d'équipements sport, loisir, culture et tourisme	Quantitative

Table 1 : Dictionnaire des variables

3 Statistiques descriptives

L'objectif de l'étude est d'expliquer la valeur foncière d'un bien lors de la mutation en fonction d'autres variables et tenter de déceler des profils de logements sur la ville de Caen. Afin de réaliser cela, une régression multiple et une analyse en composantes principales couplée à une classification sont envisagées. Cependant, avant de s'y atteler, il est nécessaire d'observer plus en détail les variables utilisées. Le but est de mieux les comprendre pour ensuite produire une analyse plus efficace. La considération des liens des variables quantitatives avec la variable de la valeur foncière et de l'influence de chaque variable qualitative est ainsi importante. Il faut de plus dénicher les potentielles interactions des variables qualitatives sur le reste des variables quantitatives et repérer de potentiels problèmes de corrélations qui pourrait altérer la régression.

3.1 Les types de logements

En moyenne, la valeur foncière est de 142 393 euros sur la ville de Caen. Près de la moitié des transactions sur la période étudiée s'estime entre 75 000 euros et 170 000 euros. Cependant, le prix le plus bas est près de 10 fois inférieur à la moyenne tandis que la transaction la plus coûteuse se dresse à plus de 10 fois la valeur moyenne. L'étendue des prix est ainsi très élevée. De plus, l'écart-type de la variable est très important, s'élevant à plus de 110 000. Ceci traduit une différence entre les prix assez importante et témoigne d'un marché foncier et immobilier très hétérogène sur la commune de Caen. Cet élément montre l'importance des différents facteurs et leur influence sur la valeur foncière.

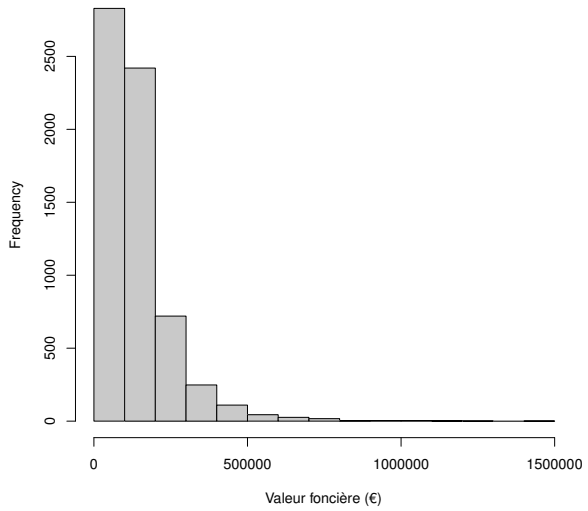

FIGURE 2 : Histogramme des valeurs foncières.

L'histogramme des valeurs foncières illustre bien le fait que la majorité des mutations s'évalue entre 0 euro et 200 000 euros. De même, la part des transactions à plus de 200 000 euros n'est pas négligeable, confirmant la dispersion importante des données.

Deux types de bien sont pris en considération : les appartements et les maisons. Parmi les données récoltées, quatre mutations sur cinq concernent un appartement. Probablement lié à cette variable, une proportion similaire de transactions accuse de la présence d'un jardin. Ces informations s'expliquent par le fait que les mutations se concentrent sur la ville de Caen ; qui dans un rapport de densité de population face à l'espace disponible comporte une majorité d'appartement sans jardin. Le recensement 2017 disponible sur le site de l'Insee indique que les appartements représentent près de 82.5% des logements à Caen.

Alors que la plus modeste surface s'élève à 10 mètres carré, le bâtiment avec la plus importante est de 467 mètres carré. La moyenne, quant à elle, se situe autour des 66 mètres carré, une valeur largement influencée par la présence majoritaire d'appartements en ville. En effet, les trois quarts des transactions sont attribuées à des bâtiments présentant une surface inférieure à 81 mètres carré. L'écart-type plutôt important se justifie par la variété binaire des types de bien.

En ne considérant que les mutations comprenant un jardin, ceux-ci mesurent en moyenne 523 mètres carré. Cependant, certains atteignent de grande étendue avec un maximum de 7 438 mètres carré. La moitié des jardins est entre 276 et 585 mètres carré, démontrant une variable assez hétérogène.

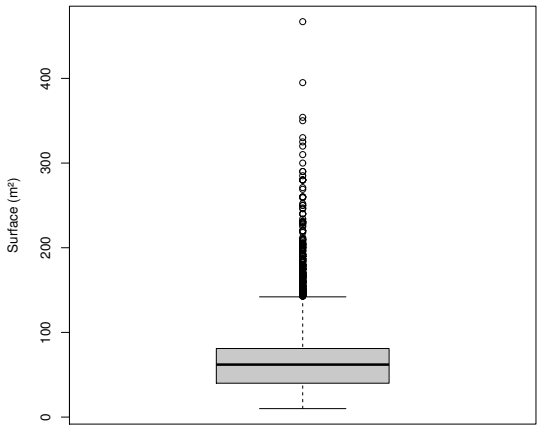

FIGURE 3 : Dispersion de la surface du bâtiment.

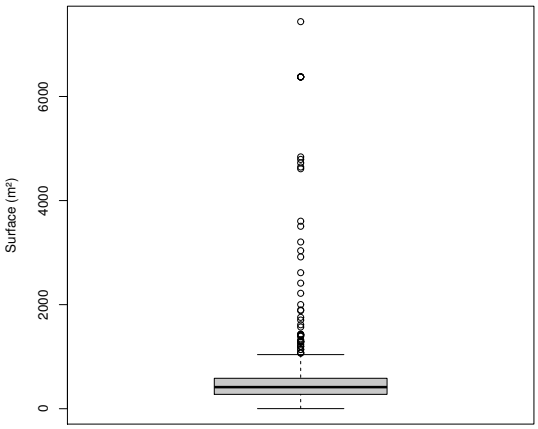

FIGURE 4 : Dispersion de la surface du terrain supérieur à 0 mètre carré.

Le boxplot de la surface du bâtiment ainsi que celui de la surface du jardin montrent de plus une importante dispersion au-dessus de la médiane.

3.2 Caractéristiques extérieures

La commune de Caen se décompose en 51 IRIS. En moyenne, les données présentent 134 mutations par IRIS sur la période étudiée sur un total de 6 429 mutations. Bien que le nombre de transactions diffère de manière plus ou moins importante entre chaque, l'écart est principalement marqué par les IRIS d'activité et divers où il est parfois observable très peu de mutations.

À Caen, le revenu déclaré médian par unité de consommation s'élève à 24 450 euros. Néanmoins, avec une distinction par IRIS, il s'étend de 8 440 euros à 29 670 euros. Ceci témoigne d'un niveau de vie disparate selon les zones qu'il pourrait être alors possible de qualifier de pauvre pour certaines et de riche pour d'autres. La variable csp représentant ici la proportion de cadres et professions intellectuelles supérieures confirme ce caractère hétérogène du niveau de vie en fonction des divers IRIS. En effet, tandis qu'un IRIS culmine à une proportion de 32%, un autre ne dépasse pas une proportion de 2%.

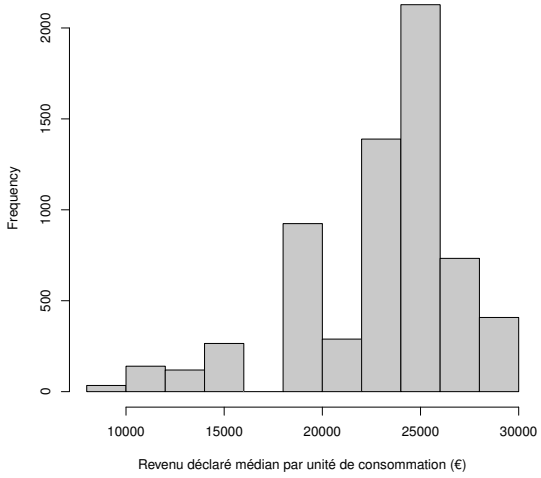

FIGURE 5 : Histogramme des revenus déclarés médians.

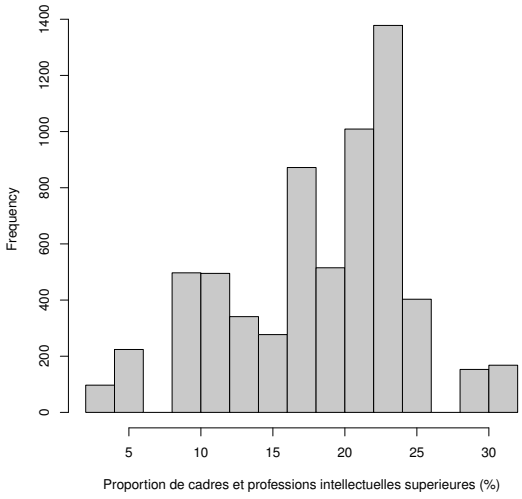

FIGURE 6 : Histogramme des proportion de cadres et professions intellectuelles supérieures.

Les histogrammes des variables rmedian et csp confirment tous les deux la dispersion importante des données. De plus, ils dénotent de manière similaire un pic au dessus de la médiane auquel s'associe assymétrie des données. Il y a plus de mutations dans les IRIS présentant un revenu déclaré médian ou une proportion de cadres et professions intellectuelles supérieures au-dessus de la médiane.

En ce qui concerne la présence d'équipement dans les IRIS de la ville de Caen, ils sont assez inégalement répartis. En effet, certains IRIS en comportent quasiment aucun alors que d'autres affichent une concentration très important avec plus de 400 équipements au total. Néanmoins, l'étude porte uniquement sur la ville de Caen. Ainsi, cette information est prévisible et compréhensible. L'utilité d'avoir un équipement précis dans chaque IRIS est nulle étant donné la proximité entre chaque IRIS.

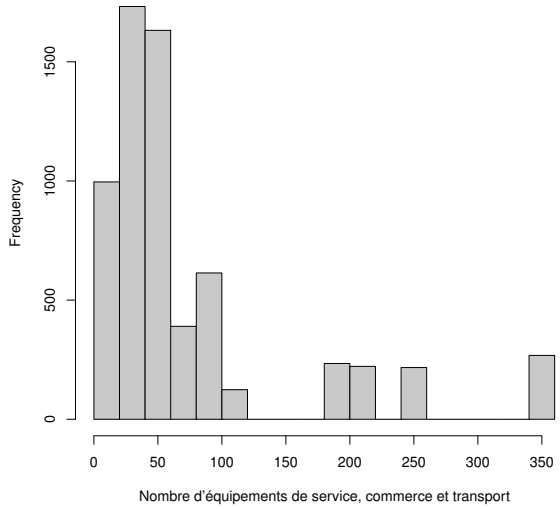

FIGURE 7 : Histogramme du nombre d'équipements de service, commerce et transport.

L'histogramme du nombre d'équipements de service, commerce et transport révèle clairement la disparité de la répartition des équipements.

Enfin, la variable zone_iris est une classification des IRIS. Le dénombrement de mutations par ces quatre zones montre une répartition disparate. En effet, le centre-ville est la zone la plus dense, accusant 28% des transactions tandis que le bord de ville l'est la moins avec seulement 14% des mutations. Les zones entre-deux et bordure-entre présentent quant à elles un nombre de ventes par IRIS proche de la moyenne.

3.3 Lien avec la valeur foncière

Dorénavant, dans l'optique d'effectuer la régression expliquant la valeur des biens, il est pertinent d'étudier le lien ou l'influence des autres variables avec cette dernière.

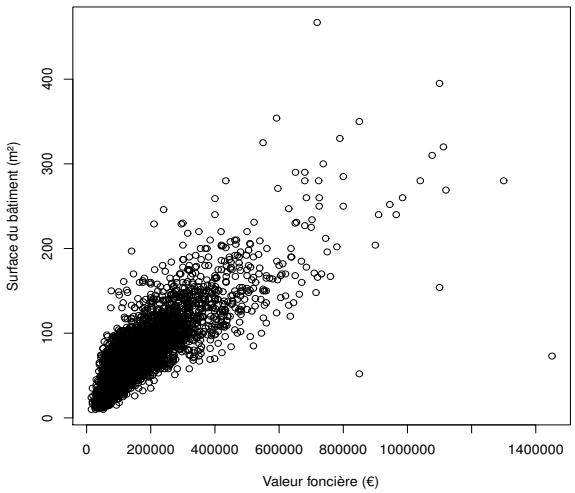

FIGURE 8 : La surface du bâtiment par rapport à la valeur foncière.

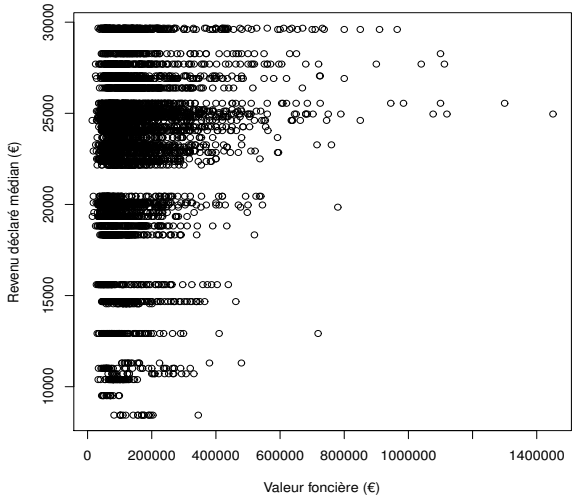

FIGURE 9 : Le revenu déclaré médian par unité de consommation par rapport à la valeur foncière.

Tout d'abord, par rapport aux variables quantitatives, uniquement la surface du bâtiment présente un lien linéaire envisageable avec la valeur foncière. Pour ce qui est du reste, non seulement un lien linéaire n'est pas envisageable, mais la nature du lien est difficillement déchiffrable. Néanmoins, ces liens existent par l'observation de variations des valeurs ou de la dispersion des données par rapport à la valeur du bien.

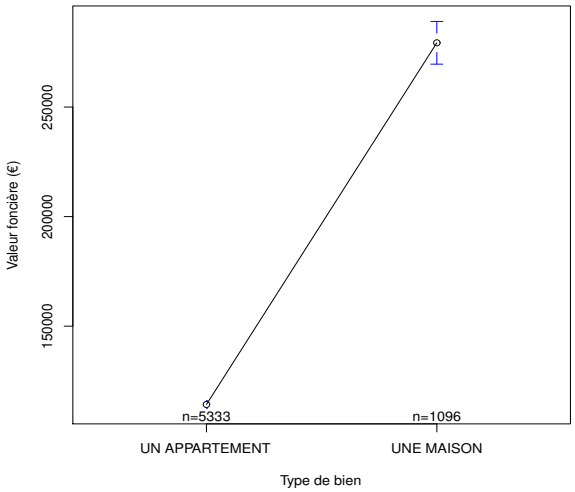

FIGURE 10 : Graphique des moyennes de la valeur foncière en fonction du type de bien.

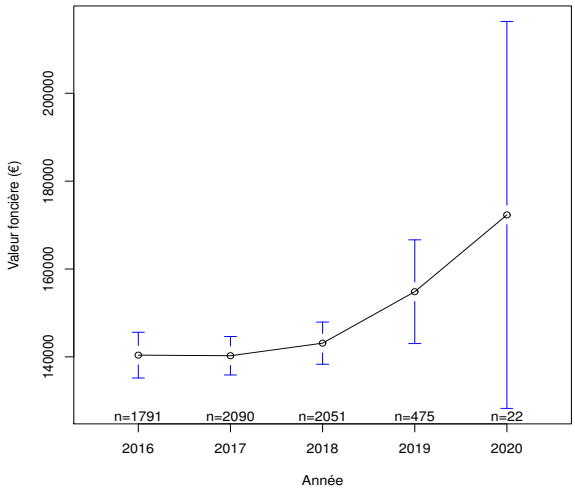

FIGURE 11 : Graphique des moyennes de la valeur foncière en fonction de l'année.

Le graphique des moyennes de la valeur foncière en fonction du type de bien illustre une différence élevée entre le prix moyen d'un appartement et celui d'une maison sur Caen. Ainsi, l'impact du type de bien sur la valeur foncière est clair.

La valeur du bien semble aussi dépendre de l'année de la mutation. En effet, le prix moyen augmente chaque année comme le montre le graphique des moyennes de la valeur foncière en fonction de l'année, faisant écho à la croissance générale de l'indice des prix des logements en France. Cette dernière information est disponible sur le site de l'Insee.

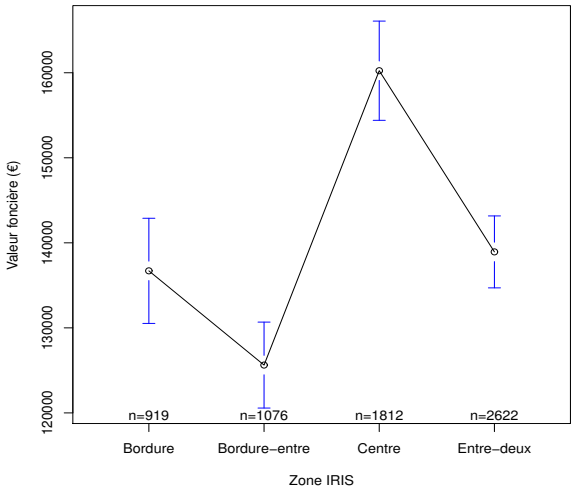

FIGURE 12 : Graphique des moyennes de la valeur foncière en fonction de la zone de l'IRIS.

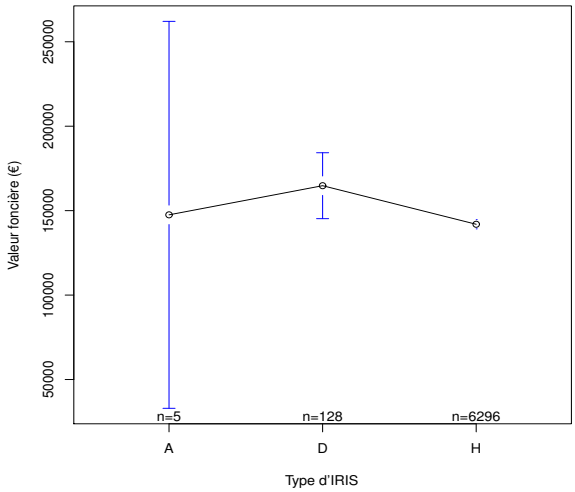

FIGURE 13 : Graphique des moyennes de la valeur foncière en fonction du type d'IRIS.

Le graphique des moyennes de la valeur foncière en fonction de la zone de l'IRIS présente l'hétérogénéité des moyennes. Il identifie une différence de plus de 30 000 euros en moyenne entre les zones "Centre" et "Bordure-entre". Cette variable qualitative a donc une influence sur la valeur du bien.

Enfin, qu'importe le type d'IRIS, la valeur foncière moyenne semble assez homogène d'après le graphique des moyennes de la valeur foncière en fonction du type d'IRIS. De ce fait, cette variable ne semble pas influencer de manière significative la valeur du bien.

Pour ce qui est des variables qualitatives, la présence ou non d'un jardin, le type de bien, l'IRIS ou bien la zone où se situe l'IRIS et dans une moindre l'année de transaction, elles ont toutes une influence sur la valeur foncière et seront donc intéressantes à prendre en compte dans la régression. À l'inverse, le type d'IRIS ne semble pas avoir de répercussions significatives.

3.4 Multicolinéarité

Afin d'écarter tout problème de multicolinéarité lors de la régression, il est nécessaire d'étudier le lien entre les variables entre elles et non seulement avec la valeur foncière.

De nature, les variables csp et rmedian représentent approximativement la même information quant au niveau de vie. Graphiquement, un lien linéaire est envisageable. Néanmoins, l'évaluation du lien linéaire entre la catégorie socioprofessionelle "cadres et professions intellectuelles supérieures" et le revenu déclaré médian par le coefficient de Sperman élevé au carré renvoie une valeur équivalente à 0.67, indiquant alors un faible lien linéaire bien que pas éloigné d'être bon. Dans ce cas, il reste pertinent de considérer les deux variables dans la régression sous la condition de contrôler la multicolinéarité.

Bien qu'étant plus ou moins liées, les trois variables portant sur l'équipement par IRIS à Caen dévoilent des liens linéaires faibles voire inexistants entre elles. Une hésitation est possible à la vue du graphique de la relation entre la variable serv_equip concernant les services et commerces d'un côté et la variable soc_equip à propos de l'enseignement et de la santé de l'autre. Cependant, l'évaluation par le coefficient de Spearman élevé au carré donne une valeur égale à 0.44, soit un lien linéaire faible. Ainsi, chacune de ces variables sont conservables pour la régression.

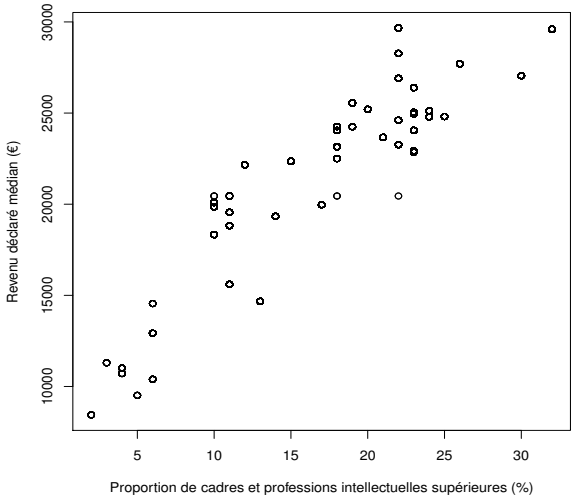

FIGURE 14 : Lien linéaire envisageable entre la proportion "cadres et professions intellectuelles supérieures" et le revenu déclaré médian .

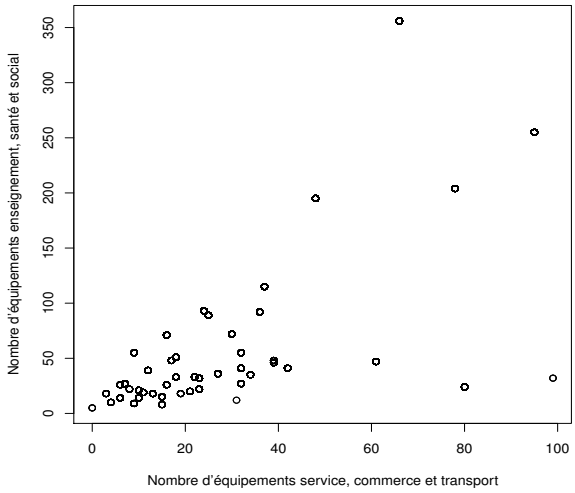

FIGURE 15 : Lien linéaire envisageable entre le nombre d'équipements de service, commerce et transport et celui d'enseignement, de santé et de social.

Les jardins étant bien souvent assimilés aux maisons et non aux appartemments, il semble évident qu'un lien entre les deux variables qualitatives existe. En effet, un calcul du V de Cramer entre les variables jardin et typebien donne une valeur de 0.94, stipulant un lien fort. Dans ce cas, il faut délaisser une des variables pour la régression. D'un côté la variable jardin traduit la présence par 1 ou l'absence par 0 d'un jardin. D'autre part, la variable sterr équivaut à la surface du terrain Étant donné la possible redondance de la variable jardin traduisant la présence d'un jardin ou non avec la variable sterr équivalente à la surface du terrain et qui, dans ce cas où les données ne concernent que des appartements et des maisons, correspond par conséquent à la surface du jardin, le meilleur choix est de conserver la variable typebien.

Avec un V de Cramer très proche de 1, le lien entre les variables code_iris et zone_iris est très fort. Cela fait sens alors que l'une découle de l'autre. Par soucis de simplicité, la variable qualitative code_iris dotée de 51 valeurs est retirée au profit du découpage en 4 zones. Le constat est similaire avec la variable nom_iris.

Toutes les relations entre variables non mentionnées présentent des liens inexistants.

3.5 Interaction

Dans l'intérêt de la régression, il est pertinent de déceler les possibles interactions des variables qualitatives sur celles quantitatives.

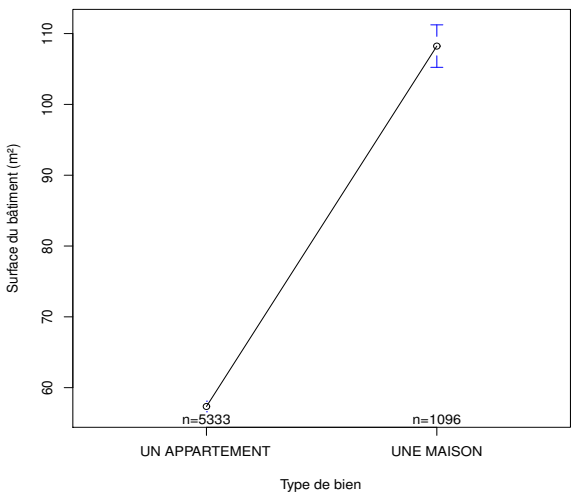

FIGURE 16 : Graphique des moyennes de la surface du bâtiment en fonction du type de bien.

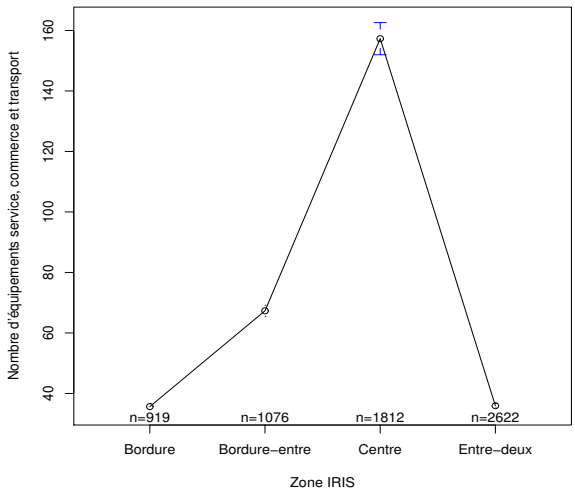

FIGURE 17 : Graphique des moyennes du nombre d'équipements de service, commerce et transport en fonction de la zone de l'IRIS.

Comme l'illustre le graphique des moyennes de la surface du bâtiment en fonction du type de bien, la surface d'une maison est significativement différente de celle d'un appartement. Le type de bien a un impact évident sur cette variable quantitative.

Le graphique des moyennes du nombre d'équipements de service, commerce et transport en fonction de la zone de l'IRIS révèle en moyenne un nombre d'équipements nettement plus élevé dans la zone "Centre". La zone où est localisé l'IRIS dénote la concentration de cette aménité urbaine.

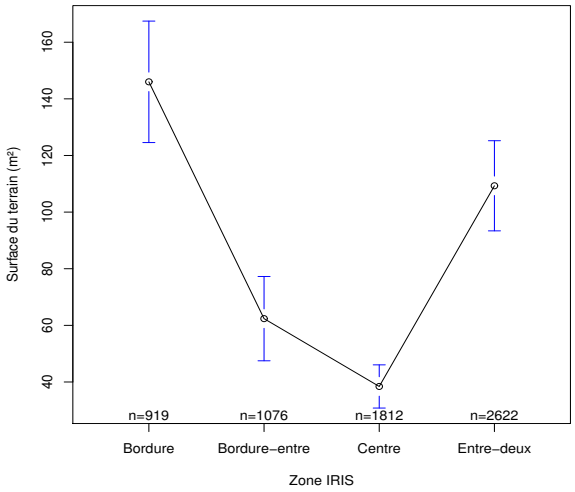

FIGURE 18 : Graphique des moyennes de la surface du terrain en fonction de la zone de l'IRIS.

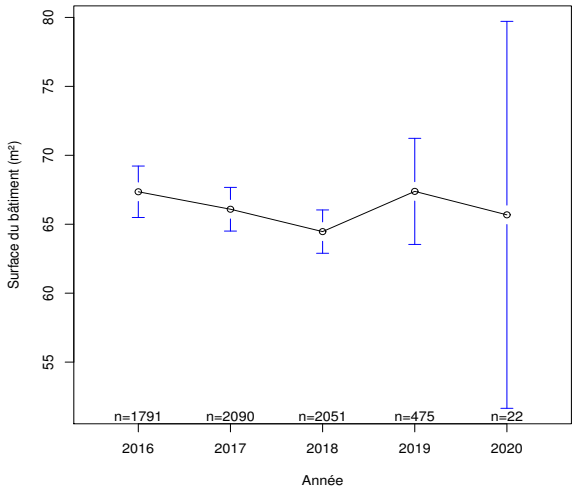

FIGURE 19 : Graphique des moyennes de la surface du bâtiment en fonction de l'année.

La disparité des moyennes de la surface du terrain est flagrante en fonction de la zone de l'IRIS comme le prouve le graphique des moyennes de la surface du terrain en fonction de la zone de l'IRIS. La localisation de l'IRIS a un impact non négligeable sur cette donnée.

Enfin, la moyenne ne fluctue pas de manière conséquente au fil de années comme le dénote le graphique des moyennes de la surface du bâtiment en fonction de l'année. L'année n'a pas d'effet considérable sur la surface moyenne du bâtiment.

Globalement, le type de bien influence la surface du bâtiment, celle du terrain, la proportion de cadres, le revenu déclaré médian ainsi le nombre d'équipements. La variable qualitative zone_iris a également un impact significatif sur chacune des variables quantitatives. Néanmoins, aucune interaction pertinente n'est révélée selon un découpage en année.

3.6 Résumé

Désormais, il est possible de lister toutes les variables utiles à la mise en place de la régression pour expliquer la valeur foncière de la transaction. En termes de variables quantitatives, il y a la surface du bâtiment et celle du terrain, les trois dénombrements d'équipements, la proportion de cadres et professions intellectuelles supérieures, le revenu déclaré médian par unité de consommation. Pour ce qui est des variables qualitatives, sont considérées le type de bien avec des interactions sur chaque variable quantitative, la zone de l'IRIS avec de même des interactions sur chaque variable quantitative et enfin l'année de la mutation.

4 Évaluation de la valeur foncière par une régression

Les variables ayant une influence plus ou moins conséquente sur la valeur foncière d'un bien sont désormais connues. L'objectif est alors d'expliquer la valeur foncière d'un bien en fonction d'informations sur la mutation et d'information sur l'IRIS. Pour y parvenir, une régréssion linéaire multiple est réalisée avec le logiciel R.

Plusieurs pistes sont envisageables quant aux résultats. Il est sensé d'anticiper que la surface du bâtiment est le critère principal dans l'estimation d'un bien. La surface du terrain a certainement une importance bien que moindre. Il peut être attendu que plus la concentration en équipements et le niveau de vie sont élevés, plus le prix sera haut par rapport à ces variables. Les différentes zones montreront sans doute des variations de l'influence de certaines variables. Ces dernières auront un effet plus important au "Centre" qu'en "Bordure" ce qui se traduira par un prix élevé au "Centre" pour des caractéristiques similaires par exemple. Une différence entre les maisons et les appartements est probable mais elle sera certainement retranscrit par la surface du terrain. Enfin, une croissance des prix des logements par année faisant écho à celle nationale est concevable. Cette régression pourra répondre à ces hypothèses.

4.1 Réalisation de la régression

Premièrement, une régression linéaire multiple impliquant chaque variable, et sans considération des interactions, est réalisée. L'intérêt est de vérifier une quelconque multicolinéarité entre certaines variables. Le modèle, sous écriture du langage R, s'exprime de la manière suivante :

$$\text{lm(valeur} \sim \text{sbati + sterr + rmedian + csp + serv\_equip + soc\_equip + slc\_equip + zone\_iris+ annee + typebien)}$$

Une vérification du facteur généralisé d'inflation de la variance de chaque variable est exécutée. Les valeurs obtenues sont contenues dans le tableau suivant :

Variable	Facteur généralisé d'inflation de la variance
sbati	1.421
sterr	1.498
rmedian	4.555
csp	4.842
serv_equip	3.817
soc_equip	2.464
slc_equip	2.214
zone_iris	2.981
annee	1.010
typebien	1.741

TABLE 2 : Tableau des facteurs généralisés d'inflation de la variance.

Toutes les valeurs sont inférieures à 5. Ainsi,il n'y a pas de problème de multicolinéarité à signaler. À titre d'indication, ce modèle présente un coefficient de détermination ajusté $\overline{R}^2$ égal à 0.74 et la p-valeur d'un test de Fisher est inférieure à 0.001. Le modèle est donc de qualité correcte en plus d'être pertinent. Néanmoins, il est souhaitable de l'améliorer.

Une seconde régression est donc effectuée. Celle-ci comporte les interactions répérées précédemment. Elle s'applique sous le modèle, sous écriture du langage R, suivant :

lm(valeur ∼ (sbati + sterr + csp + rmedian + serv_equip + soc_equip + slc_equip) * typebien + annee + (sbati + sterr + rmedian+ csp + serv_equip + soc_equip + slc_equip) * zone_iris)

Le $\overline{R}^2$ obtenu est meilleur avec comme valeur 0.782, le modèle est bon. Le test de Fisher indique un modèle pertinent et la majorité des variables ont une influence hautement significative sur la valeur foncière. Cependant, le critère d'information d'Akaike (AIC), qui est un indicateur reposant sur le compromis entre la qualité de l'ajustement et la complexité du modèle, équivaut à 157 748 ce qui est très élevé et donc de mauvaise augure. De plus, pour confirmer la validité du modèle, il est nécéssaire de vérifier les hypothèses standards. Pour ce modèle, deux problèmes sont détectés.

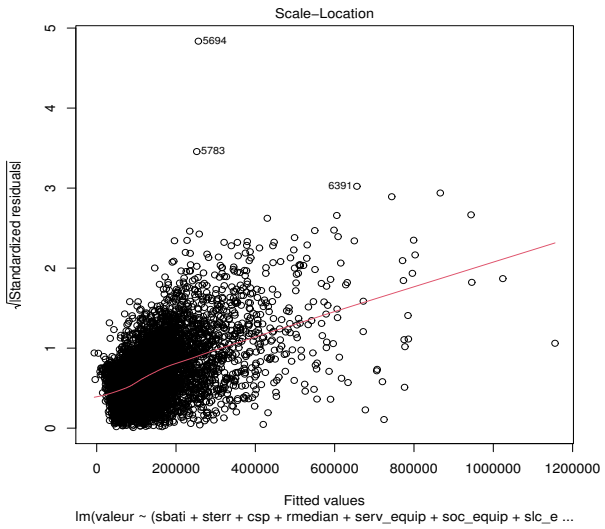

FIGURE 20 : Graphique "Scale-Location".

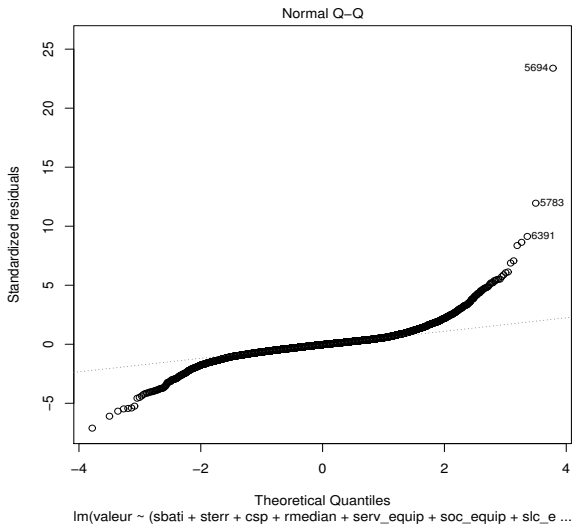

FIGURE 21 : Graphique "QQ-plot".

Le graphique "Scale-Location" montre une structure de mégaphone dans la silhouette du nuage de points. Il y a une hétéroscédasticité des variances des résidus. Sur le graphique "QQ plot", un bon nombre de points ne sont pas bien ajustés par la droite diagonale $y = x$. Dans ce cas, la normalité de la loi de chacun des résidus n'est pas admise. Ainsi, le modèle n'est pas valide.

Une solution pour pallier ces deux problèmes est la transformation de la variable expliquée, ici la valeur foncière. Une transformation avec la fonction logarithme fonctionne. Comme la variable sbati a un lien linéaire avec la valeur foncière, la fonction logarithme lui est aussi appliquée. Le nouveau modèle s'exprime, sous écriture du langage R, de la sorte :

lm(log(valeur) ∼ (log(sbati)+ sterr + csp + rmedian + serv_equip + soc_equip + slc_equip) * typebien + annee + (log(sbati) + sterr + rmedian + csp+ serv_equip + soc_equip + slc_equip) * zone_iris)

Avec un $\overline{R}^2$ de 0.797 et un AIC de 1582, le modèle obtenu est bon. Le test de Fisher indique qu'il est pertinent. Toutes les hypothèses standards semblent validées. Néanmoins, quelques variables avec une influence moindre ou peu significative sont présentes, telles que les interaction entre le type de bien et le nombre d'équipements par exemple. Il est donc judicieux de les retirer du modèle afin de gagner en simplicité avec une perte en qualité minime. Un dernier modèle est ainsi possible et sera développé de manière plus complète.

Le modèle de régression linéaire multiple retenu au final s'écrit, sous écriture du langage R, comme suivant :

lm(log(valeur) ∼ (log(sbati)+ sterr + csp + rmedian) * typebien + annee + (log(sbati) + rmedian + csp+ serv_equip + soc_equip + slc_equip) * zone_iris)

Avec le coefficient de détermination $\overline{R}^2$ et le critère d'information d'Akaike respectivement égale à 0.797 et 1591, la qualité du modèle est bonne. La p-valeur du test de Fisher est inférieure à 0.001, le modèle est donc pertinent.

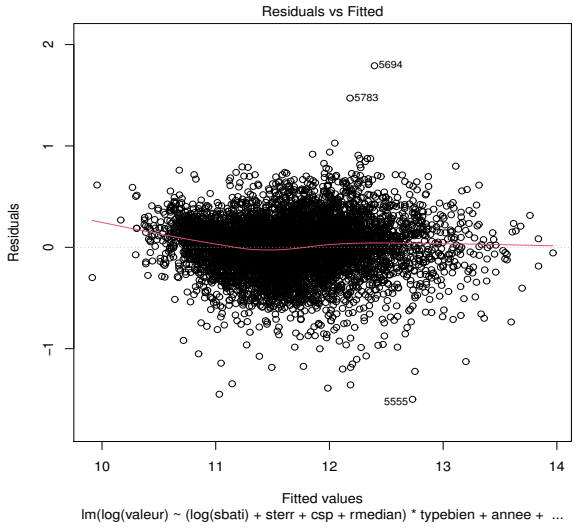

FIGURE 22 : Graphique "Residuals vs Fitted".

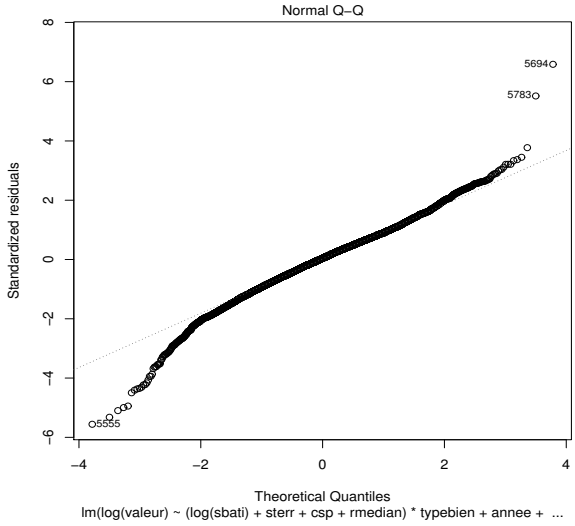

FIGURE 23 : Graphique "QQ-plot".

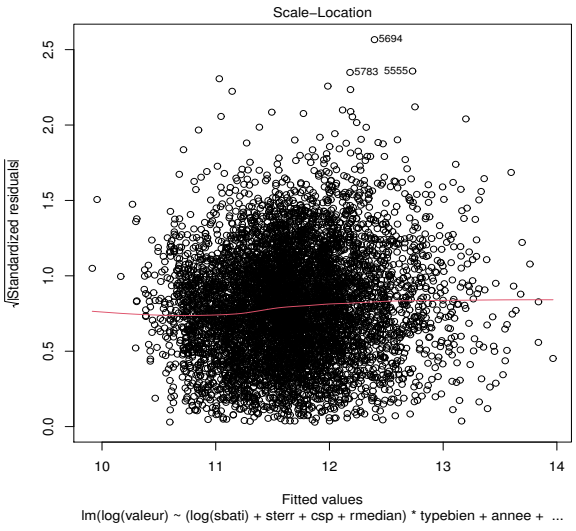

FIGURE 24 : Graphique "Scale-Location".

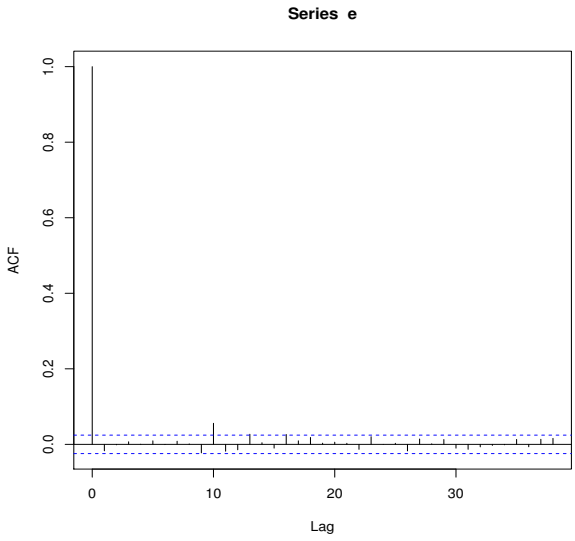

FIGURE 25 : Graphique acf des résidus.

Le graphique "Residuals vs Fitted" illustre un nuage de points difficilement ajustable par une droite, l'indépendance des résidus et des variables est admise. La normalité des résidus est acceptée par l'alignement de la majorité des points sur le graphique "QQ-plot". Il n'y a

vraisemblablement pas de structure particulière distingable depuis le nuage de points du graphique "Scale-Location". L'homoscédasticité des résidus est confirmée. Enfin, le graphique acf des résidus n'affiche pas de structure particulière et peu de bâtons dépassent les bornes limites, l'indépendance des résidus est approuvée. Ainsi, toutes les hypothèses standards sont validées, le modèle de régression est conforme.

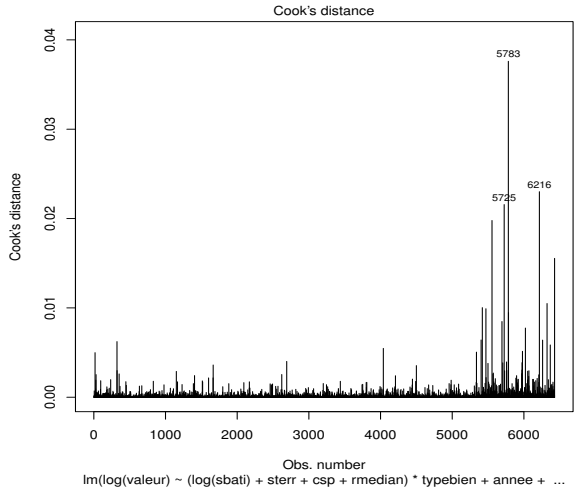

FIGURE 26 : Graphique des distance de Cook.

De plus, le graphique des distances de Cook dévoile l'absence de valeurs anormales. En effet, aucune distance ne dépasse 1.

4.2 Significativité des variables de la régression

Concernant la significativité des variables, il est à noter que toutes les variables utilisées sont d'une manière ou une autre hautement significatives, que ce soit globalement, selon une modalité ou encore via une interaction. Sur un total de 38 variables, un nombre qui inclut une variable par interaction entre deux variables, sorties par la régression, 24 affichent une p-valeur issue d'un test de Student inférieure à 0.001 et sont ainsi hautement significatives. La significativité d'une variable traduit l'influence de celle-ci sur la variable expliquée, en l'occurence la valeur foncière du bien. Une interaction ayant un haut dégré de significativité implique un impact supplémentaire d'une variable quantitative selon une modalité d'une variable qualitative. En clair, dans certaines situations, l'influence de certaines variables est d'autant plus conséquente.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	8.056e+00	1.005e-01	80.136	< 2e-16	***
log(sbati)	7.256e-01	1.967e-02	36.887	< 2e-16	***
sterr	-5.639e-05	2.823e-05	-1.998	0.045804	*
csp	-8.105e+00	9.493e-01	-8.538	< 2e-16	***
rmedian	1.005e-04	1.129e-05	8.901	< 2e-16	***
typebienUNE MAISON	6.958e-02	1.100e-01	0.632	0.527196	
annee2017	3.248e-02	8.814e-03	3.686	0.000230	***
annee2018	7.769e-02	8.863e-03	8.766	< 2e-16	***
annee2019	1.081e-01	1.415e-02	7.637	2.55e-14	***
annee2020	1.517e-01	5.869e-02	2.585	0.009754	**
serv_equip	-1.263e-02	1.850e-03	-6.829	9.32e-12	***
soc_equip	1.048e-04	1.152e-03	0.091	0.927483	
slc_equip	-7.442e-03	2.146e-03	-3.468	0.000528	***
zone_irisBordure-entre	-3.503e+00	4.306e-01	-8.134	4.97e-16	***
zone_irisCentre	6.038e-02	1.414e-01	0.427	0.669413	
zone_irisEntre-deux	3.907e-02	1.136e-01	0.344	0.730921	
log(sbati):typebienUNE MAISON	8.308e-02	2.217e-02	3.748	0.000180	***
sterr:typebienUNE MAISON	1.437e-04	3.189e-05	4.504	6.78e-06	***
csp:typebienUNE MAISON	1.375e+00	3.117e-01	4.411	1.05e-05	***
rmedian:typebienUNE MAISON	-1.286e-05	4.590e-06	-2.803	0.005086	**
log(sbati):zone_irisBordure-entre	5.758e-02	2.451e-02	2.349	0.018841	*
log(sbati):zone_irisCentre	9.485e-02	2.202e-02	4.308	1.67e-05	***
log(sbati):zone_irisEntre-deux	-1.808e-02	2.101e-02	-0.861	0.389374	
rmedian:zone_irisBordure-entre	1.561e-04	3.493e-05	4.470	7.95e-06	***
rmedian:zone_irisCentre	-7.761e-05	1.208e-05	-6.426	1.40e-10	***
rmedian:zone_irisEntre-deux	-5.911e-05	1.170e-05	-5.055	4.43e-07	***
serv_equip:zone_irisBordure-entre	1.064e-02	1.996e-03	5.333	1.00e-07	***
serv_equip:zone_irisCentre	1.313e-02	1.858e-03	7.065	1.78e-12	***
serv_equip:zone_irisEntre-deux	1.632e-02	1.901e-03	8.589	< 2e-16	***
soc_equip:zone_irisBordure-entre	-8.703e-03	1.826e-03	-4.765	1.93e-06	***
soc_equip:zone_irisCentre	-1.332e-03	1.253e-03	-1.064	0.287457	
soc_equip:zone_irisEntre-deux	-2.558e-03	1.288e-03	-1.986	0.047114	*
slc_equip:zone_irisBordure-entre	1.235e-01	1.133e-02	10.903	< 2e-16	***
slc_equip:zone_irisCentre	1.141e-02	3.984e-03	2.865	0.004185	**
slc_equip:zone_irisEntre-deux	6.086e-03	2.923e-03	2.082	0.037384	*
csp:zone_irisBordure-entre	-3.512e+00	2.160e+00	-1.626	0.104103	
csp:zone_irisCentre	6.588e+00	9.759e-01	6.751	1.60e-11	***
csp:zone_irisEntre-deux	5.547e+00	9.746e-01	5.692	1.31e-08	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

FIGURE 27 : Tableau des coefficients de la régression et leur degré de significativité.

Les codes de significativités $*$, $**$ et $***$ signifient que la variable est respectivement significative, très significative et hautement significative. Lorsqu'il n'y a rien, la variable n'est pas considérée comme significative. La variable intercept sert de base dans l'estimation et comprend en partie et en particulier les effets des modalités non présentes parmi les variables. Ici, cela correspond à une mutation datant de 2016 d'un appartement situé dans la zone "Bordure".

Les interactions s'interprètent de la manière suivante. Par exemple, l'influence du csp pour un bien en zone "Bordure" se calcule dans l'estimation telle que "(csp) $*$ (coefficient de la vairable csp)" et pour un bien localisé en zone "Centre" la formule est ""(csp) $*$ (coefficient de la vairable csp) + (csp) $*$ (coefficient de la vairable csp :zone_irisCentre)".

Plusieurs résultats confirment les hypothèses précédentes. En effet, la surface du bâtiment est bel et bien le principal critère positif dans l'évaluation de la valeur foncière. Plus le logement est grand, plus il est coûteux. Les années ont au fil du temps une influence positive de plus en plus forte sur le prix. Par exemple, un logement avec les mêmes caractérristiques sera estimé près de 16% plus cher en 2020 qu'en 2016. Une variation des effets supplémentaires

dues aux interactions est aussi observables selon les zones comme préssenti. Par exemple, pour une même surface de bâtiment, la valeur foncière sera plus élevée dans la zone d'IRIS "Centre" que dans les autres zones. Les zones "Bordure" et "Bordure-entre" sont celles où le prix du mètre carré est en général le plus bas.

L'influence du type de bien s'exprime par les interactions. Elle est effectivement retranscrite par la surface du terrain. Cette dernière variable est de base significative mais a un faible impact négatif sur le prix des appartements. Lorsque le logement est une maison, l'influence sur le prix est hautement significative positivement. Cela s'explique par le fait que moins de 1% des appartements de l'étude possèdent un terrain. En revanche, plus de 9 maisons sur 10 comportent un jardin. De plus, les résultats des interactions selon le type de bien dénotent qu'en général, le prix du mètre carré de la surface du bâtiment aura un coût supérieur pour les maisons.

Maintenant, d'autres résultats sont surprenants voire contredisent les anticipations faites. En effet, alors que le revenu déclaré médian a une influence positive et hautement significative sur la valeur foncière, la proportion de cadres et professions intellectuelles supérieures a un impact négatif et hautement significatif sur le prix. Ceci est certainement du au lien entre les deux variables, et ce même si le modèle est correct. L'aspect niveau de vie de la vairable csp est sans doute capturé par la variable rmedian. Il est alors possible qu'il y ait une potentielle compensation entre ces deux variables qui mène à une estimation plus précise de l'ensemble global nommé niveau de vie. Cette opposition reste permanente que ce soit en fonction du type de bien ou en fonction de la zone d'IRIS, bien que l'influence positive ou négative varie entre les variables. Toutefois, la combinaison des deux résulte toujours d'un niveau de vie qui a un effet positif sur le prix. Plus il est haut, plus la valeur foncière du bien sera élevée. L'impact varie selon les zones. Il sera le plus important en "Bordure-entre" et le moins au "Centre". Cela traduit certainement que le niveau de vie est plus homogène dans la zone d'IRIS "Centre" au contraire de la zone d'IRIS "Bordure-entre" où il semble être assez hétérogène.

Enfin, l'influence de la concentration en équipements est moins importante qu'attendue, du moins selon les zones d'IRIS. De plus, elle n'est pas toujours positive. Globalement, c'est surtout le nombre équipements de service, commerce et transport qui impacte le plus le prix. Sa forte concentration augmente l'estimation du prix du logement au "Centre" et en "Entre-deux". En revanche, elle cause une baisse de l'estimation dans les zones d'IRIS "Bordure" et "Bordure-entre". Le nombre d'équipements sport, loisir, culture et tourisme et celui d'enseignement, santé et social expriment leurs effets principalement dans la zone d'IRIS "Bordure-entre". L'impact est positif pour le premier ensemble et négatif pour le second. Ailleurs, l'effectif de ces équipements en particulier doit être homogène dans la zone donnée. Pour la variable slc_equip, bien qu'elle soit hautement significative de base, son plus faible impact s'explique sans doute par ses faibles valeurs. Finalement, une forte concentration d'équipements se révèle ne pas être forcément un point positif pour le prix du logement. Une proximité avec certains équipements peut être une contrainte.

4.3 Précision de la régression

Il est intéressant de considérer la précision des prédictions du modèle. Ce dernier est statistiquement valide, pertinent et de bonne qualité. Néanmoins, un coefficient de détermination ajusté égal à 0.797 montre que ce modèle n'est pas parfait et certainement améliorable. La comparaison des prédictions de la régression avec les données réelles peuvent donner une idée plus concrète de la qualité du modèle.

Par exemple, la prédiction de la valeur foncière d'une mutation datant de 2016 d'un appartement doté d'une surface de 32 mètres carré, sans terrain, situé dans un IRIS dans la zone "Centre", affichant un revenu déclaré médian de 23 270 euros et une proportion de cadres et professions intellectuelles supérieures de 22%, comptant 356 équipements de service, commerce et transport, 66 équipements d'enseignement, santé et social et 16 équipements de sport, loisir, culture et tourisme est équivalente à 82 199 euros. Dans un même temps, une transaction correspondant au même critères affiche un prix de 110 000. La prédiction est 26% plus faible que le prix réel dans ce cas.

Le prix d'une transaction réelle de la base de données datant de 2019 d'une maison ayant une surface de bâtiment de 62 mètres carré, un terrain de 456 mètres carré, situé dans un IRIS dans la zone "Entre-deux", affichant un revenu déclaré médian de 14 670 euros et une proportion de cadres et professions intellectuelles supérieures de 13%, comptant 72 équipements de service, commerce et transport, 30 équipements d'enseignement, santé et social et 2 équipements de sport, loisir, culture et tourisme est égal à 150 000 euros. Le modèle, pour des caractéristiques identiques, renvoie une valeur foncière estimée à 167 503, soit 12% supplémentaire.

Globalement, le modèle produit une prédiction de la valeur foncière à moins de 10% du prix réel pour seulement un peu moins d'une mutation sur trois. En revanche, cette proportion monte à un peu plus de trois transactions sur quatre lorsque la différence est élargie à un maximum de 30%.

4.4 Conclusion

Les variables utilisées pour l'évaluation de la valeur foncière d'un bien sont pertinentes. Néanmoins, alors que le modèle de régression linéaire multiple établi est bon, il gagnerait à être enrichi. La précision des prédictions issues de celui-ci l'illustre très bien. Il y a très certainement un manque d'information pour l'évaluation précise de la valeur foncière. L'ajout de données pourrait être bénéfique à l'étude. Par exemple, dans son article "Le prix des attributs du logement", Jean Cavailhès détermine que l'équipement sanitaire du logement et l'accès à l'emploi jouent fortement sur le prix du loyer. De même, une certaine orientation au soleil ou l'importance du vis-à-vis ont sans doute un impact dans l'estimation du prix du bien. En revanche, ces types d'informations sont difficillement ou partiellement nullement obtenable étant des informations assez détaillées. De plus, cette tentative de gain en précison et en qualité implique en parallèle un gain en complexité.

Le modèle ici étudié permet alors d'avoir une première idée de la valeur foncière d'un logement et des facteurs majeurs l'expliquant, tout en ayant une complexité raisonnable.

5 Détermination de profils de logements

Une seconde approche dans l'évaluation de la valeur foncière d'un logement est de mettre en évidence des profils de logements. Alors que la régression linéaire multiple permet de visualiser l'influence de chaque variable, une analyse en correspondances principales suivie d'une classifiaction ascendante hiérarchique donne l'opportunité d'avoir une idée de comment les différents facteurs coexistent pour expliquer le prix d'un logement.

5.1 Analyse en composantes principales

Une analyse en composantes principales (ACP) est ansi réalisée en reprenant les différentes variables quantitatives et qualitatives utilisées lors de la régression linéaire multiple.

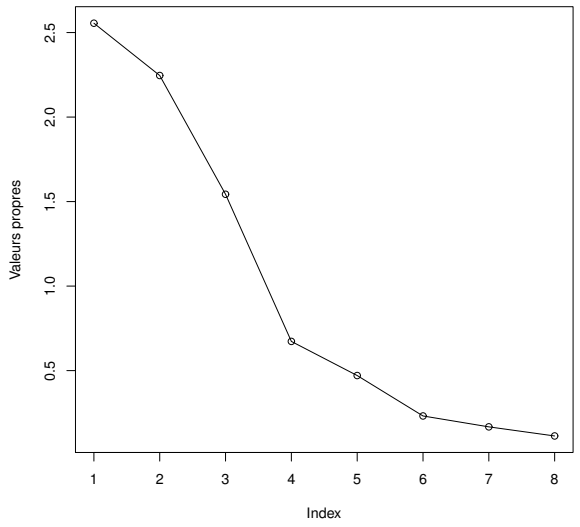

FIGURE 28 : Graphique de l'éboulis des valeurs propres.

Tout d'abord, le graphique de l'éboulis des valeurs propres montre un premier coude après la troisème valeur, indiquant la prise en compte de trois axes. L'ACP est normée et seulement les trois premiers axes ont une valeur propre supérieure à 1. Ainsi, le critère de Kaiser conduit vers trois axes également. Enfin, l'inertie expliquée des trois premiers axes est près à 80%, ce qui est suffisant. Ainsi, il semble judicieux de s'intéresser seulement aux

trois premiers axes.

Les résultats obtenus sont résumés dans le tableau suivant :

Eigenvalues

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8
Variance	2.555	2.246	1.543	0.673	0.471	0.232	0.167	0.113
% of var.	31.942	28.073	19.290	8.409	5.885	2.895	2.090	1.416
Cumulative % of var.	31.942	60.015	79.305	87.714	93.599	96.494	98.584	100.000

Individuals (the 10 first)

	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
1	4.708	3.354	0.068	0.507	-2.791	0.054	0.351	1.034	0.011	0.048
2	2.795	-1.865	0.021	0.445	-1.549	0.017	0.307	0.883	0.008	0.100
3	2.253	-2.056	0.026	0.833	-0.262	0.000	0.014	0.717	0.005	0.101
4	4.638	3.395	0.070	0.536	-2.625	0.048	0.320	1.108	0.012	0.057
5	1.836	-0.264	0.000	0.021	-0.755	0.004	0.169	-1.252	0.016	0.465
6	6.389	3.940	0.094	0.380	3.435	0.082	0.289	1.964	0.039	0.094
7	2.392	2.045	0.025	0.731	0.126	0.000	0.003	0.751	0.006	0.098
8	1.693	0.185	0.000	0.012	-1.178	0.010	0.484	-0.052	0.000	0.001
9	1.972	-0.693	0.003	0.124	0.489	0.002	0.062	-1.694	0.029	0.738
10	1.903	0.114	0.000	0.004	-1.416	0.014	0.554	-0.155	0.000	0.007

Variables

	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
valeur	0.316	3.915	0.100	0.825	30.285	0.680	0.283	5.180	0.080
sbati	0.237	2.199	0.056	0.820	29.962	0.673	0.299	5.798	0.089
sterr	0.057	0.128	0.003	0.645	18.497	0.415	0.273	4.841	0.075
rmedian	0.586	13.420	0.343	0.246	2.685	0.060	-0.731	34.651	0.535
csp	0.665	17.316	0.442	0.206	1.889	0.042	-0.671	29.148	0.450
serv_equip	0.808	25.540	0.653	-0.357	5.689	0.128	0.275	4.900	0.076
soc_equip	0.688	18.541	0.474	-0.380	6.434	0.144	0.310	6.211	0.096
slc_equip	0.696	18.940	0.484	-0.320	4.559	0.102	0.378	9.271	0.143

Supplementary categories

	Dist	Dim.1	cos2	v.test	Dim.2	cos2	v.test	Dim.3	cos2	v.test
jardin_0	0.434	0.001	0.000	0.148	-0.407	0.876	-48.783	-0.133	0.093	-19.224
jardin_1	2.185	-0.007	0.000	-0.148	2.045	0.876	48.783	0.668	0.093	19.224
UN APPARTEMENT	0.441	0.007	0.000	0.811	-0.417	0.892	-49.190	-0.133	0.091	-18.999
UNE MAISON	2.147	-0.036	0.000	-0.811	2.028	0.892	49.190	0.649	0.091	18.999
annee_2016	0.080	0.037	0.212	1.147	0.032	0.161	1.067	-0.043	0.291	-1.729
annee_2017	0.042	-0.032	0.591	-1.129	-0.003	0.005	-0.105	-0.018	0.185	-0.813
annee_2018	0.074	-0.019	0.063	-0.640	-0.044	0.345	-1.593	0.044	0.348	1.931
annee_2019	0.136	0.057	0.178	0.812	0.076	0.311	1.144	0.059	0.190	1.079
annee_2020	0.634	0.588	0.860	1.728	0.078	0.015	0.246	-0.119	0.035	-0.450
Bordure	1.513	-1.237	0.669	-25.347	0.094	0.004	2.045	0.802	0.281	21.137
Bordure-entre	0.524	0.049	0.009	1.110	-0.346	0.436	-8.306	0.167	0.102	4.841
Centre	1.657	1.589	0.920	49.933	-0.337	0.041	-11.285	-0.111	0.005	-4.498
Entre-deux	0.825	-0.685	0.689	-28.503	0.342	0.172	15.184	-0.273	0.109	-14.614

FIGURE 29 : Tableau des résultats de l'ACP.

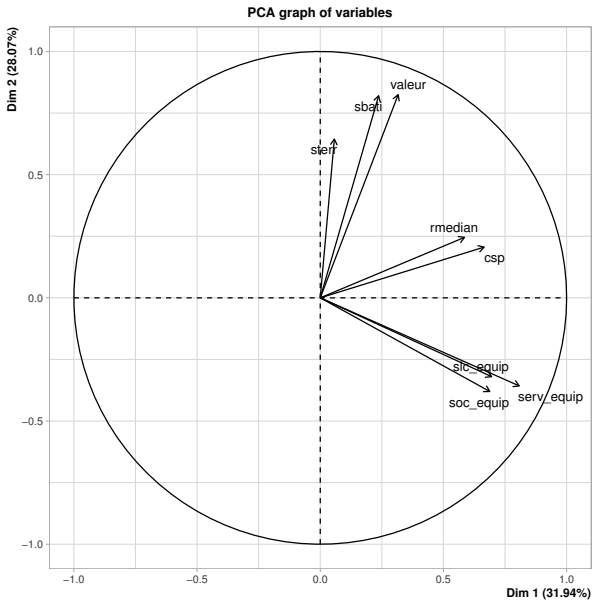

FIGURE 30 : Graphique des variables sous les axes 1 et 2.

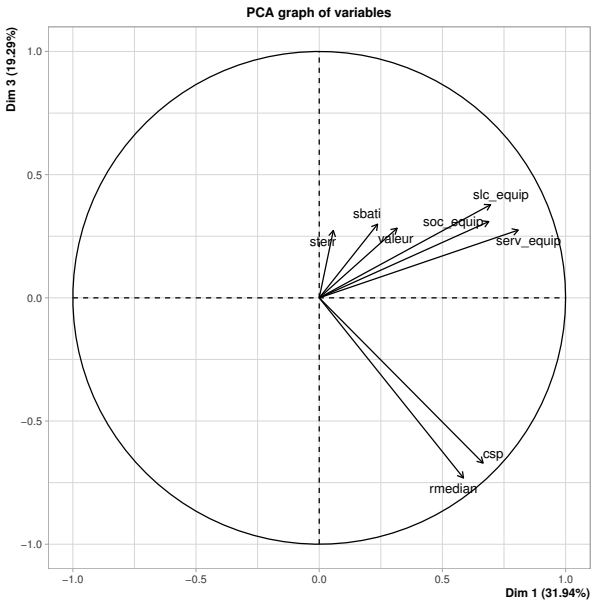

FIGURE 31 : Graphique des variables sous les axes 1 et 3.

Ce sont le nombre d'équipements en tout genre, le revenu déclaré médian et la proportion de cadres et professions intellectuelles supérieures qui contribuent le plus à l'axe 1. Le nombre d'équipements de service, commerce et transport est très bien représenté. En revanche, le reste ne l'est que moyennement. Globalement, il y a un effet taille. Certains IRIS présentent à la fois un niveau de vie et une concentration d'équipements élevés et l'inverse pour d'autres. Le découpage en zone d'IRIS est bien exposé par l'axe. Ainsi, cette dimension illustre la distinction des zones d'IRIS en fonction du nombre d'équipements principalement et du niveau de vie dans une moindre mesure. En général, les IRIS avec un haut niveau de vie et une forte concentration d'équipements se trouvent au "Centre". Les zones "Bordure" et "Entre-deux" ont globalement un niveau de vie et une concentration d'équipements plus faibles que la moyenne. En revanche, ces valeurs sont assez homogènes et proches de la moyenne dans la zone 'Bordure-entre".

L'axe 2 est marqué par la contribution de la valeur foncière, de la surface du bâtiment et de celle du terrain. Il expose le lien entre la surface et le prix. Le type de bien et la présence d'un jardin ou non s'expriment dans cette dimension. Les logements les plus chers sont en général assimilés aux maisons avec un jardin et les moins chers aux appartements sans terrain.

Le troisième axe est caractéristique du niveau de vie, bien que les variables csp et rmedian n'aient qu'une représentation moyenne. Il permet de voir que les zones ne sont pas vraiment impactées par le niveau de vie. Il est alors globalement homogène selon le découpage en zone. L'axe illustre la forte corrélation du revenu déclaré médian et de la proportion de cadres et professions intellectuelles supérieures.

5.2 Classification ascendante hiérarchique

En partant des coordonnées de chaque mutation issues de l'ACP. Une classification ascendante hiérarchique est exécutée. Quatre groupes sont établis. Ceux-ci sont composés respectivement de 2 696, 2 223, 390 et 1 120 logements, sur un total de 6 429 transactions.

```
Description of each cluster by quantitative variables
======================================================
$`1`
              v.test Mean in category Overall mean sd in category   Overall sd       p.value
soc_equip   -4.970214     3.103598e+01 3.267724e+01   1.848455e+01 2.249947e+01  6.687907e-07
sterr       -6.774448     5.439243e+01 8.671349e+01   1.579463e+02 3.250727e+02  1.248822e-11
sbati      -16.917391     5.655193e+01 6.601758e+01   2.833052e+01 3.812289e+01  3.349449e-64
valeur     -20.546658     1.091741e+05 1.423930e+05   6.463509e+04 1.101571e+05  8.244677e-94
serv_equip -27.448352     4.203338e+01 7.534998e+01   2.393623e+01 8.270146e+01 7.269477e-166
slc_equip  -30.556590     3.067136e+00 5.046508e+00   2.840783e+00 4.413583e+00 4.622361e-205
rmedian    -57.841319     1.943062e+04 2.290751e+04   3.769467e+03 4.095633e+03  0.000000e+00
csp        -63.533949     1.302485e-01 1.879095e-01   4.394634e-02 6.183650e-02  0.000000e+00

$`2`
              v.test Mean in category Overall mean sd in category   Overall sd       p.value
csp         43.711289     2.342825e-01 1.879095e-01   3.436350e-02 6.183650e-02  0.000000e+00
rmedian     40.338771     2.574197e+04 2.290751e+04   1.906030e+03 4.095633e+03  0.000000e+00
sterr       -4.786072     6.002114e+01 8.671349e+01   1.605804e+02 3.250727e+02  1.700767e-06
sbati       -5.520149     6.240711e+01 6.601758e+01   2.550588e+01 3.812289e+01  3.387120e-08
valeur      -7.966831     1.273365e+05 1.423930e+05   6.443997e+04 1.101571e+05  1.627951e-15
slc_equip  -11.877887     4.147099e+00 5.046508e+00   2.637910e+00 4.413583e+00  1.542165e-32
serv_equip -22.187098     4.386955e+01 7.534998e+01   2.226641e+01 8.270146e+01 4.575852e-109
soc_equip  -30.569232     2.087719e+01 3.267724e+01   1.008032e+01 2.249947e+01 3.139674e-205

$`3`
              v.test Mean in category Overall mean sd in category   Overall sd       p.value
valeur      51.384630     4.202095e+05 1.423930e+05   1.867754e+05 1.101571e+05  0.000000e+00
sbati       47.092902     1.541333e+02 6.601758e+01   5.352424e+01 3.812289e+01  0.000000e+00
sterr       34.417270     6.358359e+02 8.671349e+01   1.023517e+03 3.250727e+02 1.391106e-259
rmedian     14.188371     2.575962e+04 2.290751e+04   2.591467e+03 4.095633e+03  1.081412e-45
csp         12.778574     2.266923e-01 1.879095e-01   4.217282e-02 6.183650e-02  2.159926e-37
serv_equip  -7.477726     4.499744e+01 7.534998e+01   2.969171e+01 8.270146e+01  7.561939e-14
slc_equip   -9.755102     2.933333e+00 5.046508e+00   2.593022e+00 4.413583e+00  1.754244e-22
soc_equip  -10.462946     2.112308e+01 3.267724e+01   1.186830e+01 2.249947e+01  1.278188e-25

$`4`
              v.test Mean in category Overall mean sd in category   Overall sd      p.value
serv_equip 68.238258     2.286000e+02 7.534998e+01   9.024773e+01 8.270146e+01 0.000000e+00
slc_equip  60.787835     1.233214e+01 5.046508e+00   3.250610e+00 4.413583e+00 0.000000e+00
soc_equip  51.384220     6.407232e+01 3.267724e+01   2.275167e+01 2.249947e+01 0.000000e+00
csp        19.801750     2.211607e-01 1.879095e-01   2.088427e-02 6.183650e-02 2.875290e-87
rmedian    15.737485     2.465782e+04 2.290751e+04   1.262153e+03 4.095633e+03 8.370446e-56
valeur      4.381729     1.555004e+05 1.423930e+05   9.515786e+04 1.101571e+05 1.177413e-05
sterr      -6.845771     2.628214e+01 8.671349e+01   1.340957e+02 3.250727e+02 7.606515e-12
```

FIGURE 32 : Description de chaque cluster par les variables.

La colonne "v test" renseigne l'écart à la moyenne globale de la variable. Plus cette valeur absolue est grande, plus la variable est caractéristique du cluster.

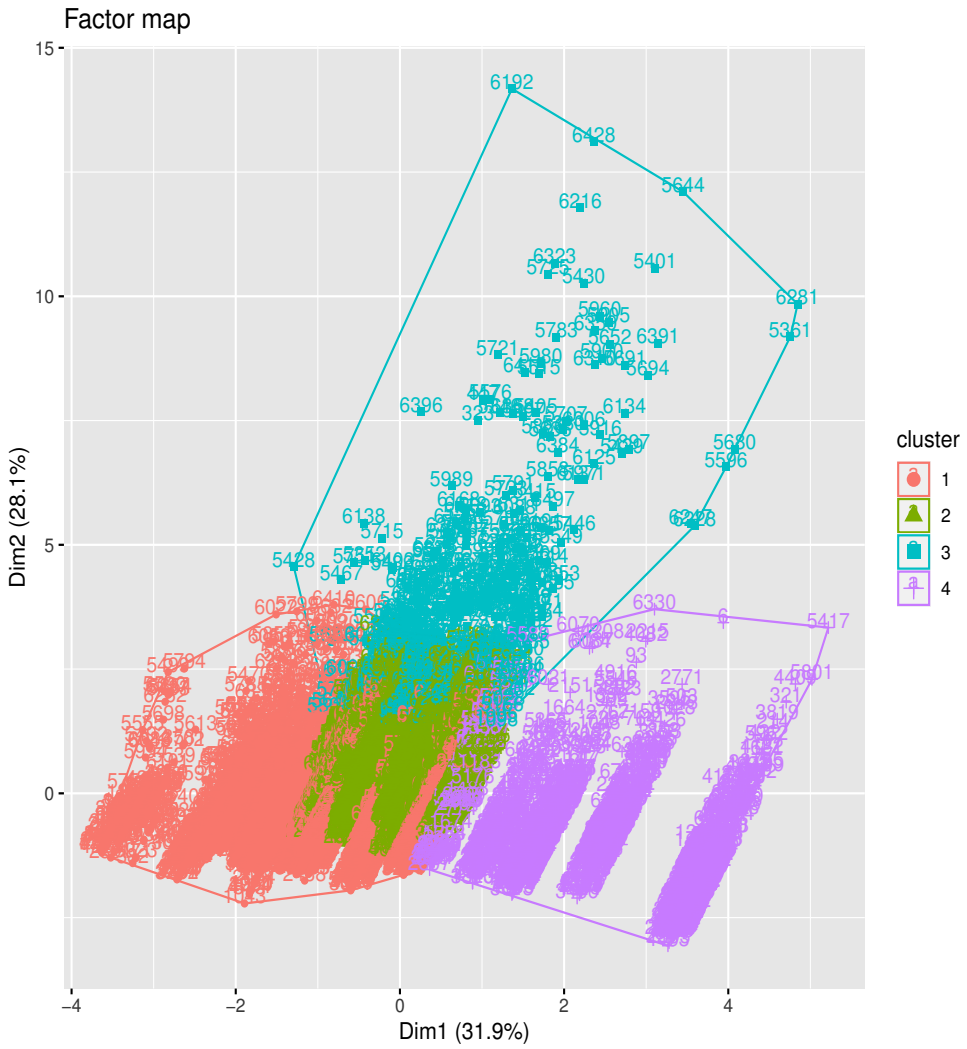

FIGURE 33 : Graphique de la classification sur les dimensions 1 et 2 de l'ACP.

Le graphique de la classification permet de visualiser clairement que les clusters 1, 2 et 4 se sont principalement construits par rapport au niveau de vie et à la concentration d'équipements. La valeur foncière et la surface du bâtiment y sont globalement autour de la moyenne communale. En revanche, il démontre que le cluster 3 s'établit surtout sur la base de la valeur foncière et de la surface du logement et celle du terrain.

D'après la description de chaque cluster par les variables, le cluster 1 est principalement marqué par un faible niveau de vie ainsi qu'une concentration d'équipements plus faible que la moyenne. De plus, il comprend en général des logements avec une valeur foncière et une surface en-dessous de la moyenne.

De son côté, le second cluster se définit par rapport à un haut niveau de vie et une concentration en équipements quelque peu plus faible que la moyenne. La valeur foncière, tout comme les surfaces, se situe aux abords de la moyenne.

Les principales caractéristiques du cluster 3 sont ses importantes surfaces de bâtiment et de terrain et surtout une valeur foncière bien au-dessus de la moyenne. Néanmoins, ce cluster ne se démarque pas par le reste des variables.

Enfin, le quatrième cluster dénote une très forte concentration d'équipements. De plus, les logements de ce cluster sont situés dans des IRIS présentant un niveau de vie légèrement plus haut que la moyenne.

Étant donné que seul le troisième cluster affiche une variation de la valeur foncière à la moyenne, il est intéressant de vérifier si il est caractéristique d'un certain profil de logements.

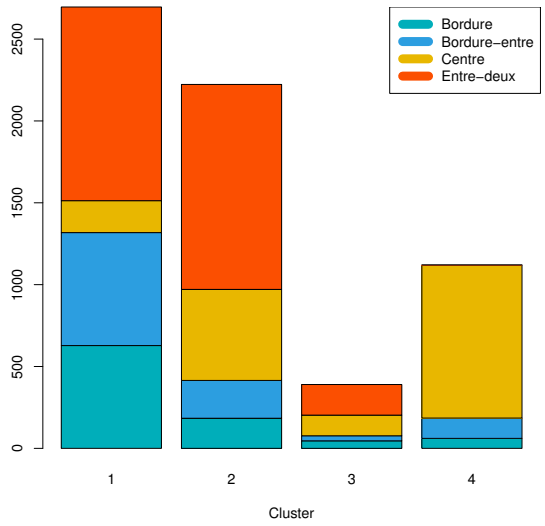

FIGURE 34 : Diagramme en bâtons des zones pour chaque cluster.

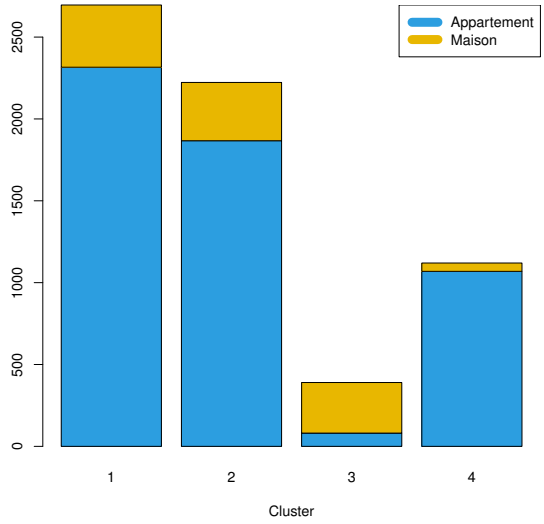

FIGURE 35 : Diagramme en bâtons du type de bien pour chaque cluster.

Le diagramme en bâtons des zones pour chaque cluster révèle que le cluster 3 est en majeure partie composé de logements localisés au "Centre" et en "Entre-deux". Cependant, l'effectif de ce cluster est assez faible. Il ne comprend que 6% et 7% du nombre total de transactions effectuées respectivement au "Centre" et en "Entre-deux". Ainsi, ce groupe n'est pas caractéristique d'une zone en particulier. Une valeur foncière assez haute n'est pas liable à un zone précise.

Le diagramme en bâtons du type de bien pour chaque cluster indique de manière similaire au précédent que le cluster 3 n'est pas réprésentatif d'un type de bien en particulier. Néanmoins, ce cluster à la valeur foncière plus haute que la moyenne est composé à près de 80% de maisons. En revanche, les autres groupes comportent seulement 5% à 15% de maisons et donc une majorité d'appartements.

5.3 Conclusion

L'objectif de cette approche est d'établir des profils de logement pour l'évaluation. Or, il y a une certaine homogénéité de la valeur foncière à Caen en termes de zones. C'est à dire qu'il n'y a vraisemblablement pas de zones où les prix sont particulièrement plus élevés qu'ailleurs en moyenne. L'analyse en correspondances principales et la classification réalisées indique que le prix dépend principalement de la surface du bâtiment et dans une moindre mesure de celle du terrain. Ainsi, les plus hautes valeurs foncières sont le plus souvent assimilées à de grand logement doté d'un grand terrain. La classification s'effectue surtout selon les caractéristiques des zones et non pas par la valeur du logement. Ces zones se différent exclusivement par la concentration d'équipements et le niveau de vie. Il n'est donc pas possible d'établir différents profils de logements avec un intérêt pour l'évaluation de la valeur foncière à Caen. Cela s'explique certainement par l'étendue des IRIS qui regroupent sous leurs caractéristiques des quartiers et des biens assez différents.

6 Conclusion de l'étude

L'utilisation des données "Demande de Valeurs Foncières" complétées par des informations sur le niveau de vie et la concentration d'équipements est pertinente dans l'évaluation de la valeur foncière d'un bien. Elle offre une première idée du prix d'un logement et des facteurs l'expliquant. Cette étude révèle que la surface du bâtiment est le facteur avec la plus forte influence dans l'explication de la valeur foncière. De plus, elle démontre de même que de nombreuses données plus discrètes rentrent en jeu dans l'évaluation du prix du logement et que, cumulées, elles ont un impact conséquent. Cependant, un enrichissement des données est nécessaire pour gagner d'avantage en précision. En revanche, bien que les découpages en IRIS apportent de nombreuses données, leur homogénéité rend la distinction de profils de logements difficile à l'échelle de la ville de Caen. L'emploi d'outils géostatistiques pourrait être intéressant pour améliorer l'étude.

Références

Adresse.data.gouv.fr (2021). Base Adresse Nationale (BAN)
https ://adresse.data.gouv.fr/donnees-nationales

Cavailhès J. (2013). Le prix des attributs du logement.
https ://www.insee.fr/fr/statistiques/fichier/1376579/es381-382e.pdf

Cerema (2021). Base de données "DVF+ open-data"
https ://datafoncier.cerema.fr/donnees/autres-donnees-foncieres/dvfplus-open-data

Data.gouv.fr (2021). Base de données "Demandes de valeurs foncières" (DVF)
https ://www.data.gouv.fr/fr/datasets/5c4ae55a634f4117716d5656/

Dvf Lifti (2019). Les données DVF.
https ://www.youtube.com/watch ?v=7kU9YBOuvo8

Garaud D. (2020). Pyris : Insee IRIS Geolocalizer
https ://pyris.datajazz.io/

Insee (2021). Dossier complet Commune de Caen (14118).
https ://www.insee.fr/fr/statistiques/2011101 ?geo=COM-14118

Insee (2021). Indice des prix des logements (neufs et anciens) – Brut – Base 100 en moyenne annuelle 2015.
https ://www.insee.fr/fr/statistiques/serie/010001868#Graphique

Insee (2016). IRIS.
https ://www.insee.fr/fr/metadonnees/definition/c1523

Insee (2020). Population en 2017. Recensement de la population - Base infracommunale (IRIS).
https ://www.insee.fr/fr/statistiques/4799309

Insee (2021). Revenus, pauvreté et niveau de vie en 2018 (Iris).
https ://www.insee.fr/fr/statistiques/5055909

Insee (2020). Dénombrement des équipements en 2019 (commerce, services, santé...).- Base permanente des équipements (BPE)
https ://www.insee.fr/fr/statistiques/3568629 ?sommaire=3568656

Travers M., Appéré G. et Larue S. (2013). Évaluation des aménités urbaines par la méthode des prix hédoniques : une application au cas de la ville d'Angers.
http ://insee.fr/fr/statistiques/fichier/1377435/ES460G.pdf