MINISTÈRE CHARGÉ DU LOGEMENT

Liberté
Égalité
Fraternité

AGRO SUP DIJON

INRAE

Les indicateurs de loyers dans le parc locatif privé

Note méthodologique

Marie Breuillé, chargée de recherche en économie, INRAE, CESAER
Camille Grivault, géographe à AgroSup Dijon, CESAER
Julie Le Gallo, Professeure d'économie à AgroSup Dijon, CESAER

Avec le soutien de Maxime Chodorge (ANIL) et Geneviève Prandi (OLAP)

01/12/2020

Table des matières

Résumé

Cette note présente la méthodologie utilisée pour estimer les indicateurs de loyers charges comprises sur la base des annonces publiées sur les plateformes leboncoin, SeLoger et pap.fr. La méthodologie repose sur la division du territoire français en mailles homogènes au vu des caractéristiques du parc locatif privé et des locataires. Au sein de chaque maille, et pour chaque segment de bien (appartements, maisons), un indicateur de loyer est prédit pour chaque commune de France, à partir des prix hédoniques estimés.

A. Introduction

Plusieurs dispositifs privés et publics d'observations des loyers du parc privé coexistent à ce jour. Les observatoires locaux des loyers (OLL) fournissent une information de bonne qualité, mais uniquement sur un territoire couvrant environ la moitié du parc locatif privé français. Pour le reste du territoire, moins urbanisé, l'information est souvent très partielle, voire indisponible. Disposer d'indicateurs de loyers fiables pour tout le territoire français −tenant compte des caractéristiques des logements et reposant sur l'exploitation d'une large base de données de loyers− représente une avancée majeure, tant pour les locataires et les propriétaires, les professionnels de l'immobilier, les investisseurs que pour les décideurs publics locaux et nationaux.

Le ministère chargé du logement[1] a initié avec l'OLAP ce projet de fournir un indicateur de loyer pour chaque commune de France. Pour ce faire, des partenariats inédits ont été noués avec les grands acteurs du numérique en matière de logement, à savoir leboncoin, SeLoger et pap.fr. Les données des annonces publiées sur ces plateformes représentent un volume considérable de données, permettant notamment d'avoir des informations sur les territoires non couverts par les OLL.

La méthodologie des indicateurs de loyers repose sur un maillage du territoire −issu d'un algorithme de clustering spatial avec contrainte de contiguité géographique− qui est d'autant plus fin (i.e., avec des mailles plus homogènes) que les données sont nombreuses. Pour chacune de ces mailles, et pour chaque type de bien (appartement, maison), un modèle hédonique est estimé, ce qui permet d'attribuer une valeur à chacune des caractéristiques des logements qui contribuent à la formation des montants des loyers. Les indicateurs de loyers proviennent de prédictions de loyers calculées pour un logement de référence. Ils correspondent à des loyers « de flux » et sont charges comprises. Ces indicateurs ont ensuite été comparés aux données des OLL et de l'entreprise PriceHubble.

Cette note méthodologique présente les données des partenaires, les traitements réalisés sur ces données puis la méthodologie retenue pour la conception des indicateurs de loyers. Enfin, les limites intrinsèques aux données et leurs implications méthodologiques dans une perspective de raffinement des indicateurs de loyers sont abordées dans la dernière partie.

B. Données des partenaires

La conception de la carte des indicateurs de loyers repose sur l'exploitation des données d'annonces de leboncoin (01/06/2016 – 31/05/2019), SeLoger (01/01/2015 – 18/02/2019) et pap.fr (01/01/2011 – 01/02/2019).

1/ Géocodage et consolidation sur la localisation géographique

La localisation des logements, variable clef pour garantir la fiabilité et la précision des estimations, n'est pas homogène entre les trois bases. La base de données leboncoin fournit

[1] Sous l'impulsion d'Eva Simon (DHUP), qui est vivement remerciée pour sa contribution.

uniquement le code postal et le nom de la commune ; la base de données pap.fr contient le code postal, le nom de la commune et parfois l'adresse du bien ; la base de données SeLoger fournit en sus le code Insee de la commune et parfois une longitude et une latitude. Toutefois, les coordonnées géographiques de la base SeLoger sont difficilement exploitables car nous ne savons pas si elles indiquent la localisation du bien ou de l'agence ayant proposé le bien à la location.

Les observations des trois bases ont été géocodées en utilisant l'API de la Banque d'adresses nationales (BAN)[2] afin de récupérer *a minima* le code Insee de la commune et si possible les coordonnées XY exactes du bien. Cette API renvoie un score permettant d'évaluer la fiabilité du résultat et différentes variables décrivant la localisation du logement : code Insee, nom de la commune, numéro et rue lorsque l'information est disponible, coordonnées géographiques, précision du géocodage. Le géocodage est considéré comme fiable uniquement lorsqu'il y a une correspondance entre les codes postaux (et les codes Insee pour SeLoger) présents dans les bases initiales et ceux obtenus par l'intermédiaire de l'application. Les observations n'étant pas géocodées à l'issue de cette première étape sont confrontées à d'autres référentiels : les fichiers de correspondance entre les codes postaux et les codes Insee de la Poste et le code officiel géographique de l'INSEE à partir du nom de la commune et du département de localisation.

2/ Consolidation des différentes variables et suppression des valeurs manquantes et extrêmes

Parmi les variables caractérisant les annonces mises en ligne sur les sites des trois partenaires (voir tableau 1), les variables communes et exploitables (car bien renseignées) sont les suivantes : le loyer charges comprises, la surface, le nombre de pièces, le type de bien, le mois et l'année de l'annonce et la commune. La décomposition entre le loyer hors charges et les charges n'est pas connue, ce qui implique d'estimer les indicateurs de loyers charges comprises.

[2] https://adresse.data.gouv.fr/api

Tableau 1 : Récapitulatif des principales variables disponibles

Variables	leboncoin	SeLoger	pap.fr
Loyer + charges	X	X	X
Surface	X	X	X
Nombre de pièces	X	X	X
Nombre de chambres	X	X	X
Type de bien (maison, appart.)	X	X	X
Étage			X
Ascenseur			X
Balcon/terrasse			X
Surface terrain			X
DPE			X
Adresse		Partiellement utilisable	X
Commune	X	X	X
Latitude/longitude		X	
Mois de location	X	X	X
Année de location	X	X	X

Chacune de ces variables fait l'objet d'un contrôle de cohérence et le cas échéant, d'un traitement adapté, pour gérer les problèmes liés à la complétude et aux incohérences (fautes de frappe, imprécisions, oublis etc.). Les observations possédant au moins une valeur manquante parmi les variables de loyer, surface, nombre de pièces, date de publication et commune de localisation, sont supprimées. Les observations présentant des valeurs aberrantes, avec un loyer, une surface ou un nombre de pièces non codés par une valeur numérique, ou (ce qui arrive peu fréquemment) une date de publication ou un code communal avec un mauvais format sont également supprimées.

Enfin, pour éviter les biais dans les estimations des loyers liés aux valeurs extrêmes, des seuils, synthétisés dans le tableau 2, sont retenus après observation des distributions. Ils concernent à la fois les variables brutes (surface, nombre de pièces, loyer) et leur ratio (surface moyenne par pièce et loyer/m²).

Tableau 2 : Seuils retenus pour supprimer les valeurs extrêmes de la base de données agrégée

Variables	Appartements	Maisons
Surface	>= 10 m² et <= 300 m²	>= 20 m² et <= 300 m²
Nombre de pièces	>= 1 pièce et <= 8 pièces (les observations pour lesquelles le nombre de pièces est égal à 0 prennent la valeur 1	>= 1 pièce et <= 13 pièces (les observations pour lesquelles le nombre de pièces est égal à 0 prennent la valeur 1
Loyer	>= 50 euros et <= 15 000 euros	>= 50 euros et <= 20 000 euros
Surface moyenne par pièce	>= 8 m² et <= 100 m²	>= 8 m² et <= 100 m²
Loyer/m²	>= 2 euros et <= 90 euros	>= 2 euros et <= 80 euros

3/ Dédoublonnage

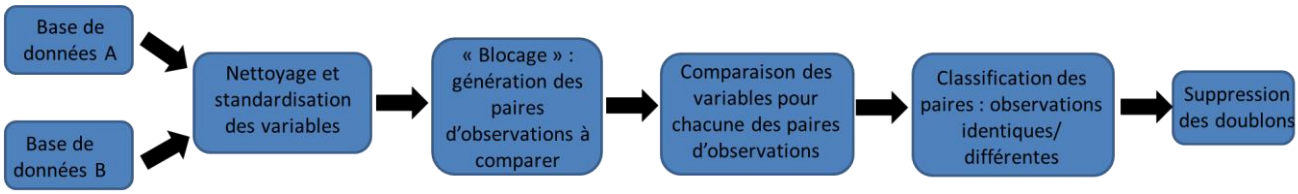
Des doublons de deux types sont susceptibles d'être présents dans les données[3] : i) des « doublons intra-compte » à l'intérieur d'une même base et pour un même compte, lorsqu'un annonceur met à jour son annonce ; ii) des « doublons inter-bases » entre les bases de données transmises par les différents partenaires, e.g., lorsque plusieurs annonces relatives à un même bien sont publiées simultanément sur différents sites.

La présence de doublons dans l'échantillon est susceptible d'introduire des biais dans les estimations des loyers, comme le montrent Sarracino et Mikucka (2017). Leurs résultats indiquent que i) si toutes les observations sont dupliquées avec la même fréquence, les estimations ne sont pas changées mais comme la taille de l'échantillon augmente, les écarts-types sont réduits ainsi que les probabilités critiques ii) si les doublons sont répartis aléatoirement dans la distribution de la variable dépendante et présents en faible proportion (moins de 10 %), ceux-ci n'ont pas d'impact sur les coefficients estimés ni sur les écarts-types iii) si les doublons ne sont pas répartis aléatoirement dans la distribution de la variable dépendante (par exemple, tous entre Q1 et Q3 ou tous avant Q1), les coefficients estimés et les écarts-types sont modifiés. Le biais augmente avec la part des observations dupliquées et les écarts-types peuvent être biaisés vers le haut ou vers le bas selon les cas. Pour limiter ces biais, la base de données constituée des annonces publiées sur les sites des trois partenaires fait l'objet d'un dédoublonnage.

a. Principales méthodes de dédoublonnage

Lorsqu'il est impossible de retrouver de façon certaine des individus (ici, des logements) en utilisant un identifiant direct, le couplage entre deux bases de données (ou le dédoublonnage d'une même base) doit se faire sur la base de l'information disponible, à condition qu'il existe des variables permettant d'identifier indirectement les individus. La démarche de dédoublonnage procède généralement en 5 étapes (Figure 1)

Figure 1 : démarche générale de dédoublonnage



La première étape consiste à nettoyer et standardiser les variables de manière à disposer d'informations comparables pour confronter les observations.

[3] Un autre type de doublons pourrait être lié au fait qu'un particulier contacte, pour louer son bien, deux agences immobilières en parallèle qui passent par la même plateforme (ou qu'un particulier crée deux comptes sur un site d'annonces en ligne, où le même bien est proposé, ce qui est peu vraisemblable). Après discussion avec les partenaires et des professionnels de l'immobilier, nous n'avons pas procédé à un dédoublonnage des annonces inter-comptes au sein d'une même base, que ce soit pour les professionnels ou les particuliers, car ce phénomène semble très marginal.

La deuxième étape consiste à générer l'ensemble des paires d'observations qui devront être comparées afin d'identifier les doublons. Cette identification est un problème quadratique qui implique de comparer ($N * N - 1$)/2 paires d'observations, où N est le nombre total d'observations. Comme il n'est pas possible de comparer l'ensemble des paires d'observations sur de très grandes bases de données, il est souvent nécessaire de réaliser un « blocage », i.e., de réduire l'espace de recherche en comparant uniquement les observations possédant un ensemble d'attributs communs. Ces attributs doivent être fiables et ne pas comporter de valeurs manquantes.

La troisième étape consiste à créer un vecteur de comparaison pour caractériser les liens entre deux observations. Plusieurs mesures de proximité sont disponibles selon la nature des données utilisées : fonction d'identité (identique/non identique), distance de Jaro-Winkler ou de Levenshtein pour les chaînes de caractère, etc.

La quatrième étape consiste à statuer sur la similarité des paires d'observations. N'importe quel classifieur binaire peut être utilisé mais quatre grandes familles de méthodes sont généralement utilisées :

- Déterministe : des règles de décision déterministes sont définies pour classer les paires ;
- Probabiliste : un score prenant en compte le pouvoir discriminant des variables est attribué et un seuil est appliqué afin d'obtenir la classification (modèle de Fellegi et Sunter (1969); Epi-Weight (Contiero et al., 2005)) ;
- Classification non supervisée : utilisation d'algorithme de type « *kmeans* » ou « *bagged clustering* » ;
- Classification supervisée : ces méthodes nécessitent la constitution d'un échantillon d'apprentissage (ex : *boosted regression trees*, *random forest*, *support vector machines*, *neural networks*, etc.).

La dernière étape consiste à identifier les groupes d'observations similaires et à sélectionner une observation à l'intérieur de ceux-ci sur la base d'une règle de décision.

b. Méthode de dédoublonnage utilisée

En l'absence d'échantillon d'apprentissage, les méthodes de classification supervisée sont exclues. Une méthode déterministe, plus simple à mettre en œuvre[4], est donc utilisée. Le dédoublonnage est réalisé en deux temps. Un dédoublonnage intra-compte, i.e., à l'intérieur d'une même base et d'un même compte d'annonceur, est tout d'abord réalisé. Les bases sont ensuite purgées des doublons inter-base. Le tableau 3 décrit en détail ces deux étapes de dédoublonnage.

[4] Le modèle probabiliste de Fellegi et Sunter a été testé mais les résultats ne sont pas probants.

Tableau 3 : Synthèse de la démarche de dédoublonnage

Type de doublons	SeLoger	leboncoin professionnel	leboncoin particulier	pap.fr
Doublons intra-compte	Les annonces à l'intérieur d'un même compte d'annonceur possèdent toutes les mêmes caractéristiques, exceptée la date de publication. Seule l'annonce la plus récente est retenue.	Deux observations constituent un doublon si elles présentent les caractéristiques suivantes : - Mêmes compte d'annonceur, code Insee de la commune, type de bien, surface et nombre de pièces - Variables « charges inclues » et « meublé » identiques ou non renseignées pour au moins l'une des deux observations - Taux d'évolution du loyer compris entre -10 et 0% - Durée écoulée entre les deux annonces <= 60 jours	Deux observations constituent un doublon si elles présentent les caractéristiques suivantes : - Mêmes compte d'annonceur, code Insee de la commune, type de bien, surface et nombre de pièces - Variables « charges inclues » et « meublé » identiques ou non renseignées pour au moins l'une des deux observations - Taux d'évolution du loyer compris entre -10 et + 10% - Durée écoulée entre les deux annonces <= 60 jours	La base de données a été livrée sans doublons à l'intérieur des comptes d'annonceurs
Doublons inter-base	La source de données comptant le plus d'observations pour une commune donnée est retenue		Couplage des deux bases. Les observations couplées sont considérées comme identiques si elles présentent les mêmes caractéristiques suivantes : - Mêmes code Insee de la commune, type de bien, surface et nombre de pièces - Taux d'évolution du loyer compris entre -10 et + 10% - Durée écoulée entre les deux annonces <= 60 jours	

Le dédoublonnage n'a été réalisé i) ni entre les annonces de professionnels, car il est peu vraisemblable qu'un bailleur passe par plusieurs agences pour mettre en location son bien, ii) ni entre les annonces de professionnels et de particuliers car le croisement des bases avec la méthode employée et compte tenu des variables disponibles devient dans ce cas très hasardeux. En outre, il est peu probable qu'un particulier mette simultanément une annonce sur leboncoin et confie un mandat à un professionnel.

c. Conséquences du dédoublonnage sur l'échantillon

Le dédoublonnage conduit à une réduction des observations beaucoup plus importante pour les appartements (*a fortiori* pour les logements de moins 3 pièces) que pour les maisons, et surtout localisée dans les grands et moyens pôles urbains. Le dédoublonnage des données SeLoger entraîne la suppression de 50,4 % des appartements et de 46,7 % des maisons. Ces pourcentages sont pour les annonces de professionnels sur leboncoin de 43,6 % pour les appartements et 29,15% pour les maisons, et pour les annonces de particuliers de 39,5% pour les appartements et 27,6% pour les maisons. La suppression plus importante d'annonces dans l'échantillon SeLoger s'explique par le dédoublonnage intra-compte plus important pour la base SeLoger. Néanmoins, le dédoublonnage des données SeLoger modifie très peu la structure de l'échantillon lors de la comparaison des principales caractéristiques des logements avant et après dédoublonnage.

Le dédoublonnage des données leboncoin est rendu plus difficile par le plus faible nombre de variables disponibles pour réaliser le dédoublonnage, notamment l'absence de l'adresse. La part de suppressions est plus importante à mesure que la population communale augmente, ce qui est probablement lié à une part non négligeable de faux positifs : 33 % des annonces de professionnels sont supprimées pour les communes de moins de 1000 habitants contre 57 % pour les communes de plus 200 000 habitants. Il en découle des différences un peu plus marquées à l'issue de la comparaison des statistiques descriptives de l'échantillon leboncoin avant et après dédoublonnage, comparativement à celui de SeLoger. Toutefois, la méthode retenue pour assembler les bases et constituer l'échantillon final permet de maîtriser en grande partie le problème. En effet, le choix de retenir pour les annonces de professionnels la base qui comporte le plus d'observations dans une commune donnée fait que les annonces de leboncoin sont la plupart du temps écartées de l'échantillon pour les grandes communes. Alors que sur France entière, la source SeLoger est retenue pour 61,4% des communes, si l'on se concentre sur les seules communes de plus de 20 000 habitants, la source SeLoger concerne 93,2% des cas. Par comparaison, pour les annonces de particuliers, les communes où sont présentes les seules annonces leboncoin particuliers sont très majoritaires (84,35%), suivie d'une association entre les données leboncoin et pap.fr (15,48%). Les communes concernées par les seules annonces pap.fr sont très minoritaires (0,17%).

4/ Echantillon de données et représentativité par rapport au parc locatif privé

A l'issue du dédoublonnage intra-compte et inter-base des bases de données des trois partenaires, l'échantillon final se compose de 9 046 540 observations, dont 7 546 945 appartements et 1 499 595 maisons, pour la période 2015-2019. Les annonces de professionnels représentent 55,56% de l'échantillon. Au total, l'échantillon est constitué à 46,4% des données leboncoin, 1,2% des données pap.fr et 52,4% des données SeLoger.

La représentativité de l'échantillon est très satisfaisante en comparaison du parc locatif privé dans sa globalité, comme en témoignent notamment les tableaux 4 et 5[5] pour la surface et la

[5] Pour précision si nécessaire, le nombre d'observations dans l'échantillon est supérieur à celui du parc locatif privé à une année donnée, car il couvre les années 2015 à 2019.

répartition dans le zonage en aires urbaines de l'Insee. En particulier, l'échantillon permet d'avoir une très bonne couverture dans les zones les plus rurales, quoique la part des observations de l'échantillon localisées dans les zones les plus rurales soit toujours légèrement plus faible que celle observée dans le parc locatif privé.

Tableau 4 : Répartition des observations selon la surface et comparaison avec le parc locatif privé de l'INSEE (2015)

Surface	Nb appart.	% appart.	Nb log. parc locatif privé	% Parc locatif privé	Nb log. parc locatif privé (emménagés récents)	% Parc locatif privé (emménagés récents)
<30.	1588477	21,05	913938	17,95	332642	23,23
30-40	1304866	17,29	900407	17,69	262962	18,36
40-60	2370709	31,41	1529823	30,05	421991	29,47
60-80	1524695	20,20	1156545	22,72	280725	19,60
80-100	487630	6,46	414653	8,15	93497	6,53
100-120	156445	2,07	116982	2,30	26908	1,88
>=120	114123	1,51	57936	1,14	13299	0,93
Total	7546945	100	5090285	100	1432024	100,00
Surface	**Nb maisons**	**% maison**	**Nb log. parc locatif privé**	**% Parc locatif privé**	**Nb log. parc locatif privé (emménagés récents)**	**% Parc locatif privé (emménagés récents)**
<30.	23317	1,55	27221	1,34	6456	1,38
30-40	39103	2,61	74587	3,67	15774	3,38
40-60	154041	10,27	252322	12,43	55733	11,95
60-80	307888	20,53	507867	25,01	112318	24,08
80-100	427262	28,49	634559	31,25	144067	30,89
100-120	258232	17,22	324629	15,99	78760	16,89
>=120	289752	19,32	209432	10,31	53278	11,42
Total	1499595	100	2030616	100	466386	100,00

Tableau 5 : Répartition des observations dans le zonage en aire urbaine

ZAU	Nb appart.	% appart.	Nb log. parc locatif privé	% Parc locatif privé	Nb log. parc locatif privé (emménagés récents)	% Parc locatif privé (emménagés récents)
Grands pôles urbains	6313432	83,66	4176656	82,05	1169645	81,68
Couronne des grands pôles	527614	6,99	356103	7,00	106037	7,40
Multipolarisé des grands pôles	141136	1,87	111090	2,18	31464	2,20
Moyens pôles urbains	189725	2,51	132739	2,61	37759	2,64
Couronne des moyens pôles	5021	0,07	4529	0,09	1322	0,09
Petits pôles urbains	195761	2,59	149235	2,93	42502	2,97
Couronne des petits pôles	2822	0,04	3022	0,06	793	0,06
Autres multipolarisés	68372	0,91	61073	1,20	17068	1,19
Rural isolé	103062	1,37	95840	1,88	25434	1,78
Total	7546945	100	5090285	100	1432024	100

ZAU	Nb maisons	% maisons	Nb log. parc locatif privé	% Parc locatif privé	Nb log. parc locatif privé (emménagés récents)	% Parc locatif privé (emménagés récents)
Grands pôles urbains	667849	44,54	817155	40,24	179687	38,53
Couronne des grands pôles	372287	24,83	487741	24,02	120028	25,74
Multipolarisé des grands pôles	107060	7,14	153040	7,54	36013	7,72
Moyens pôles urbains	63766	4,25	85007	4,19	19500	4,18
Couronne des moyens pôles	11184	0,75	17180	0,85	4118	0,88
Petits pôles urbains	83120	5,54	125036	6,16	28343	6,08
Couronne des petits pôles	3967	0,26	7959	0,39	1797	0,39
Autres multipolarisés	105658	7,05	180135	8,87	42424	9,10
Rural isolé	84704	5,65	157362	7,75	34477	7,39
Total	1499595	100	2030616	100	466386	100,00

Le tableau 6 montre cependant une surreprésentation des appartements par rapport aux maisons, surtout dans les communes peu peuplées, dans notre échantillon en comparaison du parc locatif privé.

Tableau 6 : Répartition appartements/maisons selon la population communale

Population communale	Appart. échantillon	Appart. parc locatif privé	Appart. parc locatif privé (emménagés récents)	Maisons échantillon	Maisons parc locatif privé	Maisons parc locatif privé (emménagés récents)
[0-1000]	38,88	25,99	29,49	61,12	74,01	70,51
(1000-2000]	50,66	35,94	40,19	49,34	64,06	59,81
(2000-5000]	62,74	47,37	52,31	37,26	52,63	47,69
(5000-10000]	73,73	59,40	65,12	26,27	40,60	34,88
(10000-20000]	82,69	71,30	76,03	17,31	28,70	23,97
(20000-50000]	90,81	84,21	86,80	9,19	15,79	13,20
(50000-100000]	92,47	86,80	89,23	7,53	13,20	10,77
(100000-200000]	96,45	94,36	95,60	3,55	5,64	4,40
(200000-500000]	97,15	95,35	96,51	2,85	4,65	3,49
Total	83,42	71,48	75,43	16,58	28,52	24,57

C. Maillage

Malgré la bonne couverture du territoire, pour les communes les plus rurales, le nombre d'observations dans l'échantillon est souvent insuffisant (voire parfois nul) pour procéder à des estimations à l'échelle de la commune, comme le montre le tableau 7.

Tableau 7 : Nombre de communes possédant au moins 1, 5 et 50 observations

Type de bien	Nb com avec au moins 1 obs.	Nb com avec au moins 5 obs.	Nb com avec au moins 50 obs.
Appartement	26517	17411	6204
Maison	32359	24620	6221

Le territoire est donc découpé en mailles constituées d'une ou plusieurs communes contiguës (voire d'un arrondissement pour Paris, Lyon et Marseille), à l'aide d'un algorithme de régionalisation. Le regroupement s'opère sur la base de variables qui caractérisent les logements et les locataires pour obtenir des zones où les loyers sont homogènes toutes choses égales par ailleurs.

1/ Sélection des variables caractéristiques des logements et des locataires

Les 14 variables suivantes sont retenues pour discriminer les zones où les loyers sont homogènes (tableau 8).

Tableau 8 : description des variables retenues pour la construction du maillage

Description de la variable	Source
Indice de jeunesse du parc locatif privé (Nb de logements construits après 1990 / Nb de logements construits avant 1946)	Insee, RP, 2015
Part des propriétaires occupants dans les résidences principales (%)	Insee, RP, 2015
Part des locataires du locatif privé dans les résidences principales (%)	Insee, RP, 2015
Part des locataires du locatif privé meublé dans les résidences principales (%)	Insee, RP, 2015
Part des studios et des 2 pièces dans le parc locatif privé (%)	Insee, RP, 2015
Part des résidences secondaires dans les logements ordinaires (%)	Insee, RP, 2015
Part des logements vacants dans les logements ordinaires (%)	Insee, RP, 2015
Nombre moyen de personnes par ménage dans le locatif privé	Insee, RP, 2015
Revenu médian disponible des unités de consommation (€)*	Insee, Filosofi, 2015
Taux d'évolution du nombre de ménages entre 2010 et 2015 (%)	Insee, RP, 2015
Taux de chômage (%)	Insee, RP, 2015
Part des 18-40 ans dans la population totale (%)	Insee, RP, 2015
Densité de population (nb d'habitants/ha)	Insee, populations légales, 2015
Taux de transactions dans l'ancien (Nb moyen de transactions 2010-2015/Nb de logements ordinaires en 2015)	Dv3f (2010-2016) ; Perval (2010-2016 pour l'Alsace et la Moselle) ; Insee, RP, 2015

* Cette variable n'étant pas disponible pour la Martinique et la Guyane, elle n'a pas été introduite pour réaliser leurs maillages.

Comme l'algorithme de clustering spatial utilisé n'autorise pas la présence de valeurs manquantes, une interpolation par pondération inverse à la distance est réalisée si nécessaire. Cette interpolation concerne principalement les variables « revenu médian disponible des unités de consommation » (3599 valeurs manquantes), « indice de jeunesse du parc locatif privé » (1096 valeurs manquantes), « part des studios et des 2 pièces dans le parc locatif privé » (228 valeurs manquantes) et « nombre moyen de personnes par ménage dans le parc locatif privé » (228 valeurs manquantes). Pour les autres variables, l'interpolation porte sur moins de 10 observations.

Par ailleurs, l'algorithme ne pouvant pas fonctionner en présence d'îles comptant moins de logements que le seuil retenu, les îles concernées (44 communes) sont exclues du traitement et rattachées *a posteriori* à la maille de la commune la plus proche. Lorsqu'une île est composée de plusieurs communes, toutes les communes de l'île sont rattachées à la même maille. Les communes concernées sont situées sur le littoral atlantique ainsi qu'en Martinique. Pour s'assurer que ces 14 variables décrivent bien le parc locatif privé et ses locataires, une analyse en composantes principales est menée pour résumer de la manière la plus pertinente possible les données initiales en projetant les observations dans un espace plus petit. Les 4 premières composantes principales sont retenues ; elles expliquent 54,8 % de l'inertie, c'est-à-dire de la variabilité totale du nuage des observations. Comme le montre le tableau 9, la première composante oppose les communes où la part du locatif privé est importante à celles où la part des propriétaires occupants domine. D'un point de vue géographique, elle fait apparaitre un contraste entre l'urbain d'une part et les espaces périurbains et ruraux d'autre part, ainsi qu'un contraste entre le Sud-Est du territoire, où la part du locatif privé est élevée, et le Nord-Est dominé par les propriétaires occupants. L'axe 2 est structuré par des variables caractérisant l'urbain (forte part des 18-40 ans, revenus élevés, forte croissance du nombre de ménages, faible taux de chômage, de logements vacants et de résidences secondaires). Il permet de positionner les communes le long du gradient urbain-rural. L'axe 3 oppose principalement les communes résidentielles et proches du périurbain, plutôt aisées, à des communes présentant des difficultés socio-économiques (taux de chômage élevé, part importante de logements vacants). Enfin, l'axe 4 oppose des communes en croissance démographique, touristiques, présentant un taux de chômage assez élevé et un faible taux de vacance, à celles où cette croissance est plutôt faible voire négative et le taux de vacance élevé.

Tableau 9 : Description des 4 premières composantes principales de l'ACP réalisée sur les 14 variables

Axe	Variables	Description des axes de l'ACP
Axe 1	+++ % de locatif privé +++ % des petits logements ++ % des log. meublés + Densité de population --- % des propriétaires	Opposition locatif privé/propriétaire occupant Opposition urbain/rural ; Sud-Est/Nord-Est
Axe 2	+++ % des 18-40 ans +++ Revenu médian ++ Taux d'évolution des ménages --- % de résidences secondaires -- % de logements vacants	Gradient urbain/rural

Axe 3	+++ Nb. personnes/ménage ++ Taux de chômage ++ % de logements vacants + % des 18-40 ans --- % de résidences secondaires -- % log. meublé du locatif privé -- Revenu médian - % des studios et T2 dans le locatif privé	Opposition communes proches du péri-urbain ou résidentielles / communes présentant des difficultés socio-économiques
Axe 4	+++ Taux d'évolution des ménages ++ % de résidences secondaires ++ Nb. personnes/ménage + Taux de chômage + Indice de jeunesse du parc --- % de logements vacants	Opposition communes en croissance démographique / communes moins dynamiques au fort % de logements vacants

L'ACP confirmant que les 14 variables caractéristiques retenues permettent bien d'expliquer l'hétérogénéité du parc locatif privé, elles sont introduites dans l'algorithme de régionalisation.

2/ Clustering spatial

a. Principales méthodes de clustering spatial

Les méthodes de *clustering* permettent de grouper un ensemble d'objets (ici, des communes) en des « *clusters* » (appelés « mailles »), de telle sorte que les objets se trouvant à l'intérieur d'un cluster donné soient davantage similaires en termes de caractéristiques que les objets se trouvant dans des clusters différents. La notion centrale dans ces méthodes est donc celle de degré de dissimilarité entre les objets analysés. Plus particulièrement, la méthode de clustering spatial (ou méthode de régionalisation) permet que les mailles soient composées de communes contiguës (c.-à-d. possédant une frontière commune).

Formellement, supposons n objets indicés par i, avec pour chacun p attributs (ici, les 14 variables caractérisant le parc locatif et les locataires) indicés par j. La dissimilarité entre les deux objets x_i et $x_{i'}$ est donnée comme suit :

$$D(x_i, x_{i'}) = \sum_{j=1}^{p} d_j(x_{ij}, x_{i'j})$$

avec $d_j(x_{ij}, x_{i'j})$ la dissimilarité entre les valeurs du *j-ième* attribut des objets i et i'. Le choix le plus fréquent pour cette fonction d est la distance euclidienne au carré :

$$\sum_{j=1}^{p} d_j(x_{ij}, x_{i'j}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Le choix peut également être fait de pondérer de manière différente les p attributs des objets considérés.

Lorsque les objets considérés sont des territoires géographiques, il est parfois souhaitable que ceux proches géographiquement se retrouvent dans un même « *cluster* » sans que cela ne nuise trop à la qualité du partitionnement.

L'algorithme de clustering spatial retenu ici pour partitionner le territoire français est *Max-p Region Problem* (Duque et *al.*, 2011). Cet algorithme permet de déterminer le nombre minimal d'observations (i.e., de logements) que l'on souhaite avoir dans chaque maille, et ainsi de disposer d'un nombre satisfaisant d'observations dans chacune d'entre elles pour réaliser les estimations. Par ailleurs, contrairement aux autres algorithmes, il permet d'obtenir de manière endogène le nombre final de mailles ; le maillage optimal est donc révélé par les données elles-mêmes et non par l'utilisateur. Enfin, il est relativement rapide et permet de traiter l'ensemble des quelques 36 000 communes françaises en un seul bloc, ce que ne permettent pas des algorithmes comme Skater qui nécessitent de partitionner le territoire national en sous-ensembles sur lesquels l'algorithme est implémenté (voir Colin et Roussez, 2018). Ce partitionnement est problématique car il est susceptible de créer des effets de bord à proximité des frontières utilisées pour le partitionnement.

L'algorithme *Max-p* se présente comme un programme linéaire permettant de minimiser l'hétérogénéité intra-classe sous contrainte de contiguïté. Il commence par identifier une solution faisable (un nombre *k* de clusters, contenant des communes contiguës et respectant une contrainte minimale en termes d'observations), puis itère avec pour objectif d'améliorer la solution de départ, tout en maintenant la contiguïté entre les observations de chaque cluster.

b. Mise en œuvre de l'algorithme *Max-p*

Les 14 variables caractéristiques du parc locatif privé et des locataires sont introduites dans l'algorithme de régionalisation *p-max* pour regrouper les unités spatiales présentant des caractéristiques similaires, sous contrainte de contiguïté géographique. Une contrainte additionnelle est imposée quant au nombre minimal d'observations (i.e., au nombre d'annonces) dans chaque maille, afin de garantir la qualité des estimations hédoniques. Comme ce processus est sensible aux valeurs de départ, l'algorithme a été itéré 25 fois[6] pour choisir le maillage qui minimise la variance intra-maille du loyer par m^2.

Un maillage est réalisé pour chaque type de biens (appartements, maisons). Le tableau 10 présente le nombre de mailles obtenu par type de biens, pour différents seuils d'observations (1000, 500 et 250).

Tableau 10 : Nombre de mailles en fonction de différents seuils sur le nombre d'observations et selon le type de bien

Seuil	Appartement	Maison
1000	1798	956
500	2776	1855
250	4098	3416

[6] Les temps de calculs étant importants, les 25 maillages ne sont produits que pour le seuil de 500 logements. Pour les autres seuils, un seul maillage a été produit.

Le seuil de 500 observations par maille est retenu car il constitue un bon compromis entre la finesse du maillage (et donc l'homogénéité des zones) et le nombre minimum d'observations disponibles pour les estimations réalisées au niveau de chacune des mailles.

3/ Maillage obtenu

Le maillage obtenu pour les appartements, constitué de 2 776 mailles, est le plus fin en raison du nombre d'observations plus élevé dans l'échantillon. Celui des maisons comporte 1 855 mailles.

Le maillage des appartements explique 65% de la variance du loyer/m² contre 50,4% pour les maisons. A titre de comparaison, la part de variance des loyers/m² expliquée par un maillage communal est légèrement supérieure avec respectivement 65,3% pour les appartements et 54% pour les maisons. Par ailleurs, les regroupements de communes effectués par l'algorithme de clustering n'entraînent qu'une très légère hausse de la variance intra-maille du loyer/m².

Tableau 11 : Statistiques descriptives sur le nombre de communes et d'EPCI par maille

Appartements													
Unité spatiale	min	d1	d2	d3	d4	d5	d6	d7	d8	d9	max	moy	Ecart-type
Communes	1	1	1	1	1	2	3	6	14	38	363	12,8	28,3
EPCI	1	1	1	1	1	1	1	2	3	5	19	2,3	2,5
Maisons													
Unité spatiale	min	d1	d2	d3	d4	d5	d6	d7	d8	d9	max	moy	Ecart-type
Communes	1	1	2	4	6	9	14	20	30	52	189	19,1	25,2
EPCI	1	1	1	1	2	2	3	4	5	6	16	3,1	2,4

Pour le maillage « appartements », 40 % des mailles sont composées d'une seule commune (voire un arrondissement communal pour Paris, Lyon et Marseille) et la maille la plus grande compte 363 communes. En moyenne, une maille comprend 12,8 communes et 2,3 EPCI, comme le montre le tableau 11.

Les mailles du maillage « maisons » sont plus grandes et moins de 20% des mailles ne sont composées que d'une seule commune. Elles comptent en moyenne 19,1 communes et 3,1 EPCI, et au maximum 189 communes et 16 EPCI.

Tableau 12 : Statistiques descriptives sur le nombre d'observations des mailles

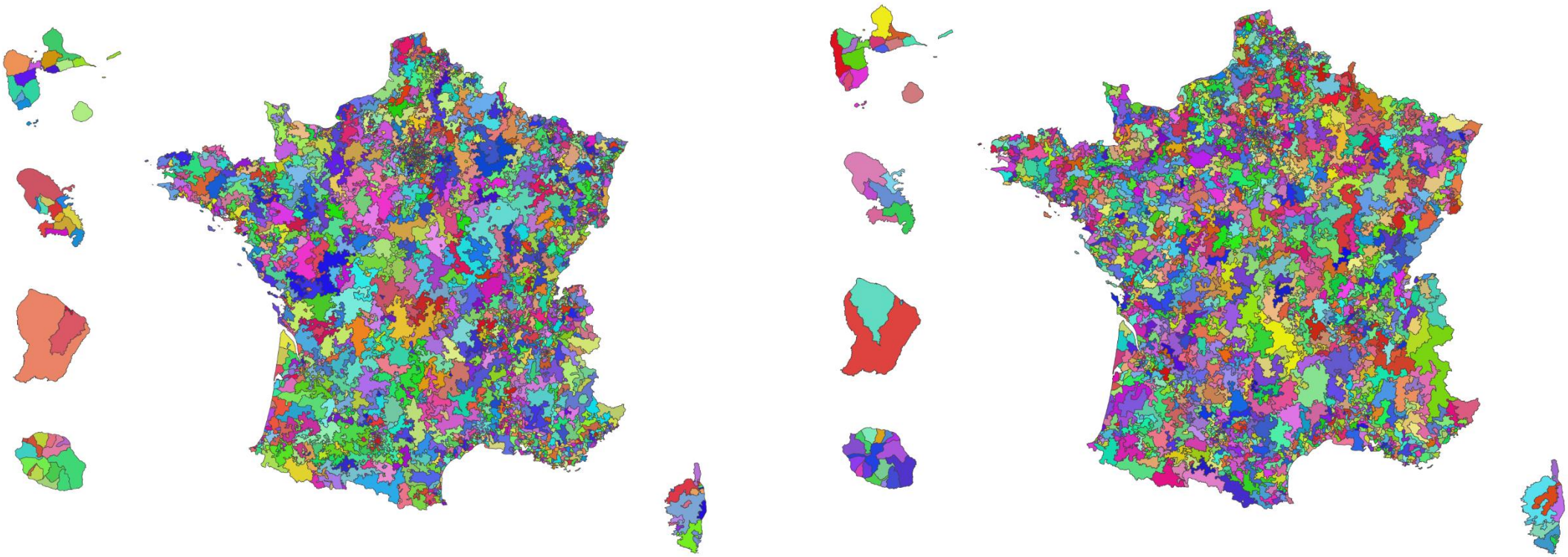
Type de biens	**Min**	**d1**	**d2**	**d3**	**d4**	**d5**	**d6**	**d7**	**d8**	**d9**	**max**	**moy**	**Ecart-type**
Appart.	500	556	626	702	794	906	1067	1376	2081	4681	210313	2719	8239
Maisons	500	536	568	607	649	698	760	836	944	1127	6887	808	417

En moyenne, une maille du maillage « appartements » compte 2 719 observations contre seulement 808 pour le maillage « maisons ». Le nombre maximal varie également fortement,

avec 210 313 observations pour le maillage « appartements » (à Toulouse) et 6 887 pour le maillage « maisons ».

Les figures 2 et 3 fournissent une représentation cartographique des maillages « appartements » et « maisons ». Comme attendu, les mailles sont très réduites dans les espaces urbains et beaucoup plus larges dans les espaces de faible densité. En effet, l'algorithme doit agrandir la maille pour atteindre le seuil des 500 observations dans les espaces de faible densité où le nombre d'observations est moindre. Les maisons étant beaucoup plus présentes dans les espaces périurbains et ruraux comparativement aux appartements qui sont plus concentrés dans les espaces urbains, les mailles pour les maisons sont de taille plus réduite dans ces espaces tandis que les mailles pour les espaces urbains sont plus larges que pour les appartements. On constate également que les maillages pour les appartements et les maisons permettent d'isoler les zones littorales qui constituent des marchés locatifs particuliers en raison de l'attractivité touristique et résidentielle.

Figures 2 et 3 : Cartographie du maillage pour les appartements (à gauche) et pour les maisons (à droite)



D. Estimation d'un modèle hédonique pour chaque maille et prédiction

Une moyenne ou médiane des loyers calculée à l'intérieur de la maille conduirait à des résultats biaisés qui ne seraient comparables entre mailles, car les caractéristiques hétérogènes des biens mis en location ne seraient pas prises en compte. Pour intégrer ces caractéristiques, un modèle hédonique est estimé à l'intérieur de chaque maille, puis un indicateur de loyer est prédit pour un bien de référence.

1/ Estimation d'un modèle hédonique de loyer pour chaque maille

Pour un maillage donné (appartements, maisons), le modèle hédonique de loyer suivant est estimé dans chacune des mailles :

$$\ln(Loyer_i) = \alpha + \beta_S f(Surface_i) + \beta_{SMP} SurfMoyPièce_i + \beta_T Trim_i + \beta_A Année_i + \beta_B Base_i + F_i + \epsilon$$

où α est la constante ; $Loyer$ est le loyer charges comprises en euros du logement i ; $f(Surface)$ est une fonction spline[7] de la surface en m² ; $SurfMoyPièce$ est la surface moyenne par pièce[8] en m² ; $Trim$ et $Année$ sont respectivement des indicatrices du trimestre et de l'année de parution de l'annonce ; $Base$ correspond à la source de données de laquelle est issue l'annonce pour le logement ; F est un effet fixe géographique et ϵ est le terme d'erreurs supposées i.i.d..

L'effet fixe est communal lorsque le nombre de communes composant la maille et possédant au moins un logement est inférieur ou égal à 50 alors qu'il est estimé à l'échelle de l'EPCI lorsque ce nombre est supérieur à 50[9]. Il n'est pas introduit lorsque la maille est composée d'une seule commune ou d'un arrondissement (pour Paris, Lyon et Marseille).

Les paramètres $\beta_S, \beta_{SMP}, \beta_T, \beta_A, \beta_B$ à estimer correspondent au vecteur des prix hédoniques et s'interprètent comme des semi-élasticités du fait de la forme semi-logarithmique.

Sous les hypothèses usuelles (entre autres $\epsilon \rightarrow N(0, \sigma^2 I)$ – voir Greene (2011) pour plus de détails), l'estimation par les Moindres Carrés Ordinaires (MCO) de ce modèle linéaire écrit sous forme matricielle :

$$Y = X\beta + \epsilon$$

où $Y = \ln(y)$ est le logarithme du loyer charges comprises, permet d'obtenir des estimateurs BLUE (Best Linear Unbiased Estimators) :

---

[7] La fonction spline est une fonction polynomiale par morceaux permettant de capter un impact non-linéaire de la surface sur le loyer en fonction de l'intervalle des valeurs prises.

[8] Comme le nombre de pièces est très corrélé avec la surface, nous avons calculé et intégré dans l'analyse la variable surface moyenne par pièce. Une spécification alternative avec prise en compte d'un potentiel effet non linéaire de la surface moyenne par pièce (i.e. une fonction spline de la surface moyenne par pièce au lieu d'une fonction spline de la surface) conduit à des R2 légèrement inférieurs et une non-significativité à 10% des coefficients pour 49 mailles appartements et 33 mailles maisons (alors que les coefficients sont toujours significatifs lorsque la fonction spline porte sur la surface).

[9] L'effet fixe est déterminé à l'échelle de l'EPCI lorsque la maille compte de nombreuses communes pour ne pas perdre trop de degrés de liberté. En effet, les mailles comportant de nombreuses communes sont également celles qui possèdent peu de logements.

$$\hat{\beta}_{ols} = (X^T X)^{-1} X^T Y$$

avec $E(\hat{\beta}_{ols}|X) = \beta$ et $V(\hat{\beta}_{ols}|X) = (X^T X)^{-1}\sigma^2$. La variance σ^2 est estimée sans biais par $\hat{\sigma}^2 = e^T e/(n-k)$ avec $e = Y - X\hat{\beta}$.

L'hétéroscédasticité est prise en compte avec une inférence robuste à l'hétéroscédasticité (estimateur de White) calculée sur des clusters issus du croisement de la variable qualitative relative à la source de données et à celle relative à l'effet fixe géographique.

2/ Identification et suppression des valeurs atypiques

Les estimations sont réalisées en deux étapes, pour obtenir des résultats plus fiables. La première étape consiste à identifier les observations atypiques, c'est-à-dire celles pour lesquelles la valeur estimée s'écarte trop de la valeur réelle et auxquelles l'équation hédonique s'applique mal. Sont considérées comme atypiques les observations dont la valeur estimée par le modèle s'écarte de la valeur réelle de plus de deux écarts-types. Les résidus standardisés sont calculés comme suit :

$$\widehat{r_{m,\iota}} = \frac{\widehat{\epsilon_{m,\iota}}}{\hat{\sigma}_m \cdot \sqrt{1 - h_{m,ii}}}$$

avec :

$\hat{\sigma}_m$: la racine carrée de la variance estimée du résidu $\widehat{\epsilon_{m,\iota}}$ égale à

$$\hat{\sigma}_m^2 = \frac{\sum_1^{n_m} \widehat{\epsilon_{m,\iota}}}{n_m - (p_m + 1)}$$

n_m : le nombre d'observations dans la maille

p_m : le nombre total de variables (caractéristiques du logement, source, indicatrices d'année ou de trimestre) dans le modèle associé à la maille

$h_{m,ii} = \chi_{m,i}{}'(X_m{}'X_m)^{-1}\chi_{m,i}$: l'effet levier de l'observation *i* où X_m est la matrice de taille $n_m * (p_m + 1)$ représentant les valeurs des variables du modèle (ainsi que la constante) pour l'ensemble des observations de la maille *m* et où $\chi_{m,i}$ désigne le vecteur de taille $1 * (p_m + 1)$ regroupant les valeurs des variables pour l'observation *i* de la maille *m*.

Il s'avère que la part des observations atypiques dans l'échantillon oscille entre 4,02% pour les appartements et 4,41% pour les maisons.

La seconde étape consiste à relancer les estimations sur l'échantillon, en excluant toutes les observations avec un résidu standardisé en dehors de l'intervalle]-2 ; 2[, à l'instar de la méthodologie des indices notaires-Insee des prix des logements anciens.

La distribution des R2 ajustés des modèles estimés en excluant les valeurs atypiques est plutôt satisfaisante, comme le montre le tableau 13. Les R2 moyens s'élèvent à 0,77 pour les

appartements et 0,73 pour les maisons. Les R2 minimums sont très faibles pour les appartements (0,18) mais plus élevés pour les maisons (0,33). Ces faibles R2 se situent surtout dans les zones rurales pour lesquelles l'algorithme a dû fortement agrandir la maille en ajoutant suffisamment de communes afin d'atteindre le seuil de 500 observations. Les R2 augmentent cependant très rapidement comme en témoigne la valeur du premier décile.

Tableau 13 : Statistiques descriptives sur les R2 ajustés obtenus

	Appartement	Maison
Min.	0,1763	0,3258
10%	0,6627	0,6198
20%	0,7138	0,6626
30%	0,7434	0,6898
40%	0,7656	0,7105
50%	0,7852	0,7328
60%	0,8033	0,7570
70%	0,8219	0,7794
80%	0,8465	0,8032
90%	0,8712	0,8386
Max.	0,9998	0,9695
Moy.	0,7728	0,7311
Ecart-type	0,0927	0,0844

3/ Prédiction

La valeur Y d'un logement avec les caractéristiques X_{new} est prédite comme suit :

$$\hat{Y}_{new} = X_{new}\,\hat{\beta}$$

où X_{new} est le vecteur contenant les valeurs du bien de référence pour lequel on souhaite obtenir une prédiction. Dans le calcul de la variance de cette prédiction, il faut non seulement prendre en compte l'incertitude liée à $\hat{\beta}$, mais également l'incertitude liée à ϵ :

$$Var(\hat{Y}_{new}) = Var(X_{new}\,\hat{\beta}) + Var(\epsilon) = X_{new}^{T}(X^{T}X)^{-1}X_{new}\sigma^{2} + \sigma^{2}$$

et l'intervalle de prédiction[10] suivant :

$$\hat{Y}_{new} +/- t_{n-k}^{1-\alpha/2}\,\hat{\sigma}\,(1 + X_{new}^{T}(X^{T}X)^{-1}X_{new})$$

Pour la spécification considérée, l'indicateur de loyer (c'est-à-dire le loyer prédit) dans la maille *i* est calculé comme suit pour le bien de référence :

[10] L'intervalle de prédiction est plus large que l'intervalle de confiance, reflétant l'incertitude supplémentaire liée à une prédiction hors échantillon. En effet, l'intervalle de confiance est l'intervalle de variation de E(Y/X) alors que l'intervalle de prédiction est l'intervalle de variation de Y.

$$indicateur\ de\ loyer_i = \hat{\alpha} + \widehat{\beta_S} f\left(\overline{Surface}\right) + \widehat{\beta_{SMP}}\left(\overline{SurfMoyPIèce}\right) + \widehat{\beta_T}\ 3èmeTrim + \widehat{\beta_A}\ 2018 + \widehat{\beta_B}\ SeLoger/lbcpro + \widehat{F_i}$$

où $\widehat{\beta_S}$, $\widehat{\beta_{SMP}}$, $\widehat{\beta_T}$, $\widehat{\beta_A}$, $\widehat{\beta_B}$ et $\widehat{F_i}$ sont les coefficients estimés[11].

Les caractéristiques du bien de référence retenu pour chaque type de bien sont présentées dans le tableau 14.

Tableau 14 : Caractéristiques retenues pour réaliser les prédictions de loyers

Type de biens	**Appartement**	**Maison**
Surf.	49 m²	92 m²
Surf. moy./pièce	22,1 m²	22,5 m²
Trim.	3ème trim.	
Année	2018	
Source	- leboncoin pro. lorsque la modalité SeLoger n'est pas disponible - leboncoin pro. lorsque le coefficient associé à la modalité SeLoger n'est pas compris entre le 5ème et le 95ème centiles et lorsque la valeur absolue du coefficient associé à leboncoin pro. est inférieure à celle du coefficient SeLoger - SeLoger dans les autres cas.	
Effets fixes	Moyenne des effets fixes (commune ou EPCI) pour les communes ou EPCI comptant moins de 100 logements. Effets fixes (communes ou EPCI) dans le cas contraire, dès lors qu'il y a plus d'une commune dans la maille.	

Pour les variables quantitatives (surface et surface moyenne par pièce), nous retenons la moyenne de la variable sur l'ensemble de l'échantillon (i.e., toutes mailles confondues) pour le type de bien considéré, soit par exemple une surface de 49 m^2 et une surface moyenne par pièce de 22,1 m^2 pour les appartements. Les prédictions sont réalisées pour un logement proposé à la location au troisième trimestre de l'année 2018. La modalité source de référence est fonction de la règle suivante :

- Si $\widehat{\beta_B} SeLoger$ est compris entre le 5ème et le 95ème centiles, la source SeLoger est utilisée. En dehors de cet intervalle, le coefficient $\widehat{\beta_B} lbcpro$ est utilisé, sauf s'il est plus extrême.

- Si la source SeLoger n'est pas disponible, le coefficient $\widehat{\beta_B} lbcpro$ est utilisé.

- Dans les autres cas, le coefficient $\widehat{\beta_B} SeLoger$ s'applique.

Enfin, l'effet fixe géographique communal (resp. EPCI, quand la maille est composée de plus de 50 communes) correspond à la moyenne des effets fixes communes (resp. EPCI) de la maille lorsque la commune (resp. l'EPCI) compte moins de 100 observations. Dans les autres cas, l'effet fixe géographique est celui de la commune (resp. l'EPCI, quand la maille est composée

[11] Les indicateurs sont obtenus en appliquant une fonction exponentielle sur le logarithme des loyers prédits. Appliquer un facteur d'ajustement comme le propose Duan (1983) n'a qu'un impact très marginal (de l'ordre de la 2ème décimale pour le loyer au m^2) qui concerne davantage les loyers maximaux.

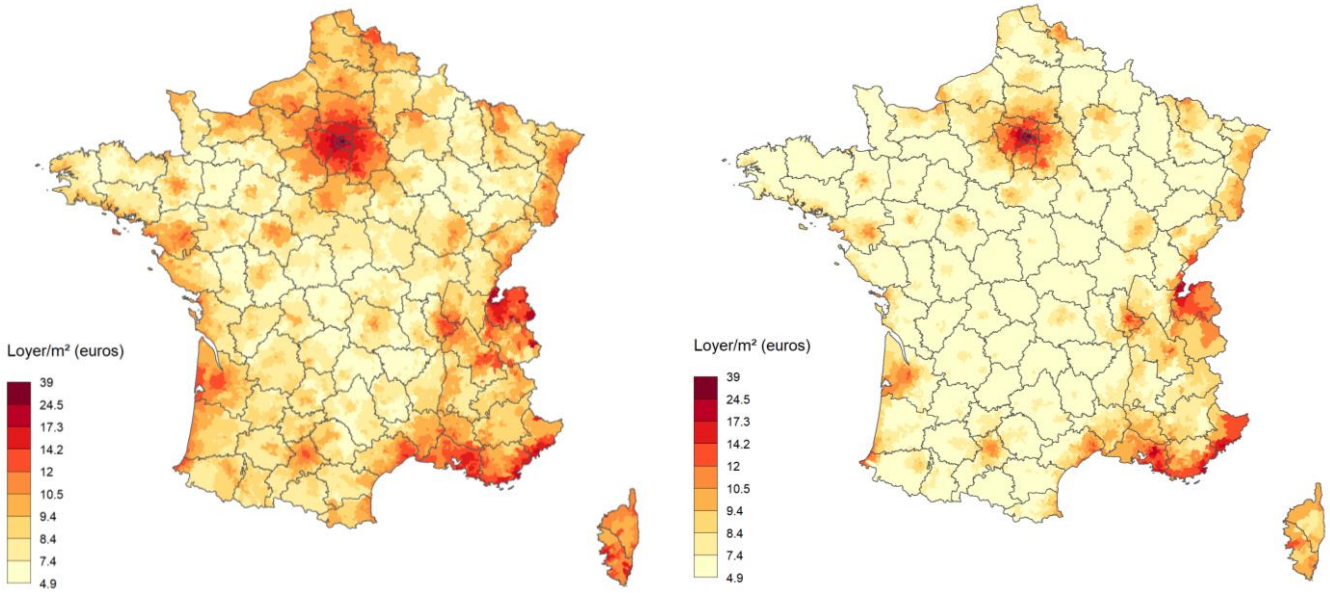
de plus de 50 communes), dès lors que la maille contient plusieurs communes. Cette procédure permet de faire varier le loyer prédit à l'intérieur de certaines mailles.

La distribution des indicateurs de loyers obtenus pour ces biens de référence sont présentés dans le tableau 15, et leur répartition géographique est cartographiée par les figures 4 et 5.

Tableau 15 : Distribution des loyers et des loyers/m² prédits

	Loyer		Loyer/m²	
	Appart.	Maison	Appart.	Maison
Min.	263	451	5,38	4,90
10%	350	528	7,15	5,74
20%	369	562	7,52	6,11
30%	386	591	7,87	6,42
40%	403	620	8,22	6,74
50%	422	646	8,60	7,02
60%	440	678	8,98	7,37
70%	464	723	9,47	7,86
80%	496	778	10,11	8,45
90%	551	876	11,24	9,52
Max.	1796	3504	36,65	38,09
Moy.	441	686	9,01	7,46
Ecart-type	99	178	2,02	1,93

Figures 4 et 5 : Cartographie des loyers/m² prédits pour les appartements (à gauche) et pour les maisons (à droite)



E. Analyse des limites et éléments conclusifs

Les indicateurs de loyers obtenus présentent certaines limites relatives aux données utilisées et à la méthode employée.

La première limite concerne l'utilisation d'annonces. Le loyer affiché sur l'annonce peut différer du loyer inscrit dans le bail de location (correspondant au loyer « de marché »). La différence est néanmoins limitée par le fait que seul le loyer de l'annonce la plus récente est retenu. Si les différences entre le montant affiché dans l'annonce et celui effectivement payé peuvent être importantes pour les transactions immobilières, l'écart est toutefois moindre voire négligeable s'agissant des mises en location réalisées par des professionnels. Il peut cependant être réel pour des mises en location par des particuliers, surtout dans des secteurs ruraux. L'utilisation d'annonces ne permet pas non plus de connaître la valeur du loyer « de stock ».

La seconde limite porte sur l'impossibilité de distinguer les charges. L'échantillon ne permet pas de distinguer le loyer et les provisions pour charges locatives. Un doute subsiste également sur le fait que tous les loyers présents dans la base de données SeLoger soient bien des loyers charges comprises. Des indicateurs de loyers hors charges pourraient être construits à condition que les partenaires soient en mesure de distinguer le loyer net des charges locatives.

La troisième limite concerne les variables structurelles et de localisation. Les variables disponibles sont peu nombreuses et limitées aux caractéristiques structurelles principales des logements. Des variables importantes telles que la période de construction, l'étage, la présence d'un ascenseur, les éléments de confort présents dans le logement, l'étiquette DPE ou encore la surface du terrain pour les maisons ne sont pas disponibles, ou le sont dans les champs non structurés de pap.fr, alors qu'elles constituent des déterminants importants des loyers. Ceci entraîne une perte de pouvoir explicatif du modèle et potentiellement un biais de variables omises si une variable explicative est corrélée avec une variable omise. Pour remédier à ce problème, il serait nécessaire d'avoir plus de caractéristiques relatives aux logements, par exemple grâce aux textes associés aux annonces. La période de construction des logements pourrait également être obtenue à partir d'un croisement avec les fichiers fonciers à condition de disposer de l'adresse précise des logements. Par ailleurs, l'absence de localisation à l'adresse pour l'ensemble de l'échantillon ne permet pas de prendre finement en compte l'hétérogénéité spatiale inobservée et de construire des variables additionnelles décrivant l'environnement du logement (socio-économique, l'accessibilité aux services ou aux aménités environnementales) ce qui entraîne un biais de variables omises. Enfin, l'autocorrélation spatiale entre les logements ne peut être prise en compte (les erreurs sont ici supposées indépendantes et seule une inférence robuste par rapport à l'hétéroscédasticité est mise en œuvre).

La quatrième limite concerne la suppression des doublons à l'intérieur et entre les différentes bases de données qui pourrait être réalisée beaucoup plus finement grâce à l'analyse textuelle et à l'exploitation des photos dans l'annonce, avec des techniques de machine learning, nécessitant des infrastructures et moyens humains conséquents.

La cinquième limite concerne l'identification des meublés. Les données SeLoger ne permettent pas d'identifier les locations de meublés, contrairement aux autres bases. Des estimations en incluant la variable meublée (avec comme référence la modalité non déclaré) conduisent

cependant à des résultats peu modifiés, sauf pour les loyers maximaux qui deviennent très importants. De même, il se peut que des biens qui ne relèvent pas du secteur locatif privé (par des bailleurs sociaux, voire par des constructeurs de maisons) soit à la marge mis en location via les sites en ligne.

Une dernière limite provient de l'endogénéité probable de la variable « base » relative à la source de données. En effet, il n'est pas certain que les biens appartenant aux différentes bases, et notamment à celles de PAP et de SeLoger, présentent des caractéristiques similaires dans chacune des mailles de la zone d'étude ce qui est susceptible d'engendrer un biais de sélection.

Bibliographie

Cailly, Cédric, et al. (2019). Les indices Notaires-Insee des prix des logements anciens Méthodologie v4, *Insee Méthodes* n°132.

Colin, S. & Roussez, V. (2018). Elaboration d'une maille géographique pour l'habitat, 13ièmes Journées de méthodologie statistique de l'Insee, 12-14 juin 2018, Paris.

Contiero et al., (2005). The EpiLink record linkage software. *Methods of information in medicine*, 44(01), 66-71.

Duan, N. (1983). Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association*, 78(383), 605-610.

Duque, J. C., Anselin, L., & Rey, S. J. (2012). The max-p-regions problem. *Journal of Regional Science*, 52(3), 397-419.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210.

Greene, W. H. (2011). Economic analysis. Pearson Education.

Sarracino, F., & Mikucka, M. (2017). Bias and efficiency loss in regression estimates due to duplicated observations: a Monte Carlo simulation. *Survey Research Methods*, 11(1), 17-44.