MINISTÈRE CHARGÉ DU LOGEMENT
Liberté
Égalité
Fraternité

anil
Agence Nationale pour l'Information sur le Logement

Les indicateurs de loyers dans le parc locatif privé

Note méthodologique

Mise à jour par l'ANIL en décembre 2022 de la note produite en décembre 2020 par :
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

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 de leboncoin et du Groupe SeLoger. 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, appartements 1-2 pièces, appartements 3 pièces et plus, maisons), un indicateur de loyer est prédit pour chaque commune de France, à partir des prix hédoniques estimés.

A. Introduction

Depuis sa création par la loi ALUR en 2014, le réseau des Observatoires locaux des loyers (OLL) a permis d'améliorer la connaissance des loyers du parc privé. Un peu plus de la moitié du parc locatif privé français est aujourd'hui couvert par un réseau de 32 OLL, qui publie chaque année des informations sur les loyers pratiqués dans une cinquantaine d'agglomérations, regroupant des communes majoritairement inscrites dans des zones tendues. Pour le reste des communes, notamment celles situées en zones non tendues, l'information est disparate, parfois partielle, voire inexistante.

Pour construire des indicateurs complémentaires à ceux produits par les OLL, homogènes et comparables au plan national, le Ministère a lancé en 2018 le projet de « carte des loyers », en s'associant à une équipe de recherche en économie d'Agrosup Dijon et de l'INRAE afin de définir une méthodologie. Celle-ci convoque des techniques économétriques et prend appui sur un partenariat inédit avec le Groupe SeLoger, leboncoin, et pap.fr. Une première carte a été ainsi construite à partir des données d'annonces publiées sur ces plateformes et a été mise en ligne en décembre 2020 sur le site du Ministère.

Le Ministère a confié à l'Agence nationale pour l'information sur le logement (ANIL) le portage de la carte de loyers, qui consiste en son actualisation régulière intégrant de nouvelles avancées méthodologiques.

Grâce à des partenariats reconduits avec leboncoin et le Groupe SeLoger pour la transmission de données d'annonces locatives, l'ANIL a pu éditer une nouvelle carte des loyers sur la base de la méthode développée par l'équipe de recherche et en apportant quelques évolutions qui sont précisées dans ce document.

La méthodologie mise en œuvre pour construire des indicateurs de loyers repose sur un maillage du territoire − issu d'un algorithme de clustering spatial avec contrainte de contiguïté 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 (appartements, appartements 1-2 pièces, appartements 3 pièces et plus, maisons), 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 au loyer du marché de la relocation et sont charges comprises.

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 en 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 et du Groupe SeLoger publiées sur leurs plateformes respectives entre le 01/01/2018 et le 30/09/2022. Sur cette période, le volume initial d'annonces réceptionnées par l'ANIL est de 24 millions. Après différents retraitements (détaillés ci-après), 7 millions de données sont finalement exploitées pour produire la carte des loyers.

1/ Géocodage et consolidation sur la localisation géographique

La localisation des logements est une variable clef pour garantir la fiabilité et la précision des estimations. L'information géographique contenue dans les bases de données porte sur la commune (ou l'arrondissement pour Paris, Lyon et Marseille). La base de données leboncoin fournit le code postal et le nom de la commune ; la base de données SeLoger fournit en sus le code Insee de la commune.

Les observations des deux bases ont été géocodées en utilisant l'API de la Banque d'adresses nationales (BAN)[1] afin de récupérer (ou vérifier) le code Insee de la commune. Cette API renvoie un score permettant d'évaluer la fiabilité du résultat et différentes variables décrivant la localisation du logement : code Insee et nom de la commune. 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.

Cependant, à l'issue de cette étape, il subsiste des observations pour lesquelles la localisation est manquante ou considérée comme non fiable (22 232 données pour SeLoger et 62 908 données pour leboncoin).

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 deux 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 (ou l'arrondissement). La décomposition entre le loyer hors charges et les charges n'est pas connue pour la base leboncoin, ce qui implique d'estimer les indicateurs de loyers charges comprises. Le caractère chargé ou non du loyer est bien renseigné dans les deux bases SeLoger et leboncoin, ce qui permet d'exclure les quelques loyers non chargés de la base SeLoger.

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

Le caractère meublé ou non du logement est renseigné dans les deux bases SeLoger et leboncoin.

Par ailleurs, en l'état actuel des données transmises, toutes les bases n'offrent pas de variables permettant de faire la distinction entre location saisonnière et location longue durée.

Enfin, il convient de mentionner l'absence de variables communes aux deux bases importantes pour expliquer les loyers telles que l'année ou la période de construction, l'étage, l'étiquette énergétique, la présence d'équipements ou d'éléments de confort du logement (nombre de toilettes et de salles de bain, présence d'un balcon (ou d'une terrasse), d'un ascenseur, d'un gardien, d'une place de parking) ou encore la surface du terrain pour les maisons.

Après différents traitements visant à consolider les informations disponibles dans les bases de données et à les rendre cohérentes en vue de leur fusion, la base de données globale ainsi constituée fournit un échantillon conséquent et relativement fiable pour estimer des modèles de loyer.

Tableau 1 : Récapitulatif des principales variables disponibles

Variables	**Leboncoin**	**SeLoger**
Loyer + charges	X	X
Charges comprises ou non dans le loyer	X	X
Montant des charges (si connues)		X
Surface	X	X
Nombre de pièces	X	X
Nombre de chambres		X
Type de bien (maison, appart.)	X	X
Étage		X
Garage/parking		X
Balcon/terrasse		X
Surface terrain		X
DPE/GES		X
Année de construction		X
Nom de la commune	X	X
Code postal	X	X
Code Insee		X
Date de mise en ligne de l'annonce	X	X
Nombre de photos dans l'annonce		X
Caractère meublé ou non du logement	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

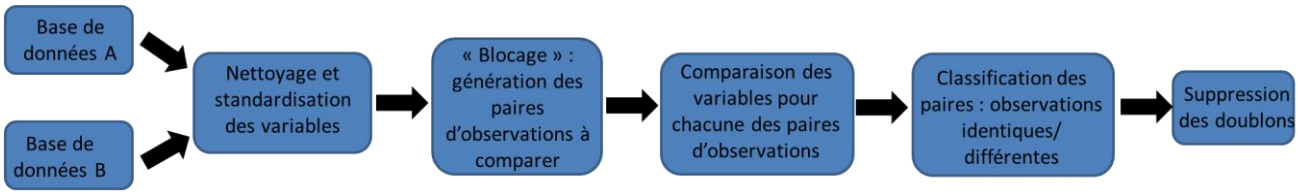
Des doublons de trois types sont susceptibles d'être présents dans les données : i) des « doublons intra-compte » à l'intérieur d'une même base et pour un même compte, lorsqu'un annonceur met à jour son annonce ; ii) des « doublons inter-compte », à l'intérieur d'une même base et pour un compte annonceur différent, iii) 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 deux 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.

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[2], est donc utilisée. Le dédoublonnage est réalisé en trois temps :

- On procède au dédoublonnage intra-compte, à l'intérieur d'une même base et d'un même compte d'annonceur ;
- Sur la base de données purgée des doublons intra-compte, on procède au dédoublonnage inter-compte (à l'intérieur d'une même base pour des comptes d'annonceurs différents) ;
- Enfin, sur les bases purgées des doublons intra et inter-compte pour les professionnels ou uniquement intra-compte pour les particuliers, on procède au dédoublonnage inter-base des annonces.

Tableau 3 : Synthèse de la démarche de dédoublonnage

Type de doublons	SeLoger	leboncoin professionnel	leboncoin particulier
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.	Les annonces à l'intérieur d'un même compte d'annonceur ne possèdent pas systématiquement toutes les mêmes caractéristiques. Aussi, 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%	Les annonces à l'intérieur d'un même compte d'annonceur ne possèdent pas systématiquement toutes les mêmes caractéristiques. Aussi, 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%

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

		- Durée écoulée entre les deux annonces <= 60 jours	- Durée écoulée entre les deux annonces <= 60 jours
Doublons inter-compte	Deux observations constituent un doublon si elles présentent les caractéristiques suivantes : - Même code Insee - Variables nombre de chambres et accès handicapé identiques ou non renseignées pour au moins l'une des deux observations - Distance de Jaro-Winkler >= 0.9 sur le libellé de l'adresse et de l'annonce ou variables 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 : - Même code Insee de la commune, type de bien, type d'annonceur, 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	Aucun dédoublonnage puisqu'il n'y a aucune raison de penser que plusieurs particuliers différents déposent une annonce pour le même bien
Doublons inter-base	La source de données comptant le plus d'observations pour une commune donnée est retenue		

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 28,2 % des appartements et de 12,2 % des maisons. Ces pourcentages sont pour les annonces de professionnels sur leboncoin de 41,2 % pour les appartements et 7,1% pour les maisons, et pour les annonces de particuliers de 40,8% pour les appartements et 11% pour les maisons.

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 :

- Pour la base leboncoin, 35 % des annonces de professionnels sont supprimées pour les communes de moins de 1000 habitants contre 70 % pour les communes de plus 200 000 habitants
- Pour la base SeLoger, 13 % des annonces de professionnels sont supprimées pour les communes de moins de 1000 habitants contre 48 % pour les communes de plus 200 000 habitants

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

À l'issue du dédoublonnage intra-compte et inter-base des bases de données des deux partenaires, l'échantillon final se compose de 7 182 710 observations, dont 5 750 382 appartements et 1 432 328 maisons, pour la période 2018-2022. Les annonces de professionnels représentent 49 % de l'échantillon. Au total, l'échantillon est constitué à 86 % des données leboncoin et 14 % 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[3] pour la surface et la population des communes. 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 (2019)

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.	899 310	15,64	632 988	13,51	186 130	16,20
30-40	915 095	15,91	829 576	17,70	211 798	18,43
40-60	1 921 380	33,41	1 525 181	32,55	380 216	33,09
60-80	1 344 048	23,37	1 148 997	24,52	257 416	22,40
80-100	452 520	7,87	393 310	8,39	80 754	7,03
100-120	138 331	2,41	106 749	2,28	22 469	1,96
>=120	79 698	1,39	49 166	1,05	10 414	0,91
Total	**5 750 382**	**100,00**	**4 685 967**	**100,00**	**1 149 197**	**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.	9 718	0,68	21 532	1,09	4 301	1,04
30-40	26 392	1,84	65 002	3,30	12 330	2,98
40-60	140 319	9,80	241 219	12,26	48 775	11,80
60-80	309 431	21,60	503 281	25,59	102 924	24,90
80-100	439 925	30,71	625 406	31,80	130 164	31,50
100-120	255 149	17,81	314 552	15,99	69 380	16,79
>=120	251 394	17,55	195 820	9,96	45 401	10,99
Total	**1 432 328**	**100,00**	**1 966 812**	**100,00**	**413 275**	**100,00**

[3] Pour précision si nécessaire, le nombre d'observations dans l'échantillon peut être supérieur à celui du parc locatif privé à une année donnée, car il couvre les années 2018 à 2022.

Tableau 5 : 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]	2,74 %	5,22 %	5,06 %	17,01 %	24,36 %	23,84 %
[1000-2000]	3,68 %	5,84 %	5,99 %	14,09 %	17,28 %	17,72 %
[2000-5000]	9,41 %	14,14 %	14,68 %	22,18 %	25,81 %	26,39 %
[5000-10000]	10,77 %	15,96 %	16,50 %	15,66 %	17,90 %	17,89 %
[10000-20000]	12,29 %	9,13 %	9,50 %	10,79 %	5,63 %	5,61 %
[20000-50000]	21,37 %	16,68 %	15,89 %	10,72 %	4,83 %	4,65 %
[50000-100000]	12,96 %	11,46 %	10,34 %	4,93 %	2,44 %	2,20 %
[100000-200000]	17,72 %	13,65 %	13,28 %	3,39 %	1,24 %	1,19 %
[200000 et +]	9,05 %	7,92 %	8,78 %	1,22 %	0,51 %	0,50 %
Total	100,00 %	100,00 %	100,00 %	100,00 %	100,00 %	100,00 %

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 6[4].

Tableau 6 : 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 10 obs.	Nb com avec au moins 50 obs.
Appartement	25 766	13 223	6 498
Maison	31 670	17 938	6 248

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.

[4] Le territoire français est composé de 34 980 communes (arrondissements pour Paris, Lyon et Marseille) au 1er janvier 2022 (hors Mayotte).

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 7).

Tableau 7 : 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, 2019
Part des propriétaires occupants dans les résidences principales (%)	Insee, RP, 2019
Part des locataires du locatif privé dans les résidences principales (%)	Insee, RP, 2019
Part des locataires du locatif privé meublé dans les résidences principales (%)	Insee, RP, 2019
Part des studios et des 2 pièces dans le parc locatif privé (%)	Insee, RP, 2019
Part des résidences secondaires dans les logements ordinaires (%)	Insee, RP, 2019
Part des logements vacants dans les logements ordinaires (%)	Insee, RP, 2019
Nombre moyen de personnes par ménage dans le locatif privé	Insee, RP, 2019
Revenu médian disponible des unités de consommation (€)*	Insee, Filosofi, 2019
Taux d'évolution du nombre de ménages entre 2014 et 2019 (%)	Insee, RP, 2019
Taux de chômage (%)	Insee, RP, 2019
Part des 18-40 ans dans la population totale (%)	Insee, RP, 2019
Densité de population (nb d'habitants/ha)	Insee, populations légales, 2019
Taux de transactions dans l'ancien (Nb moyen de transactions 2017-2021/Nb de logements ordinaires en 2019)	Dv3f (2021) ; Perval (2016-2018 pour l'Alsace et la Moselle) ; Insee, RP, 2019

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

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 » (3 632 valeurs manquantes), « indice de jeunesse du parc locatif privé » (1 228 valeurs manquantes), « part des studios et des 2 pièces dans le parc locatif privé » (221 valeurs manquantes) et « nombre moyen de personnes par ménage dans le parc locatif privé » (221 valeurs manquantes). Pour les autres variables, l'interpolation porte sur moins de 29 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 55,1 % de l'inertie, c'est-à-dire de la variabilité totale du nuage des observations. Comme le montre le tableau 8, 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 8 : 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 -- % Taux de chômage	Gradient urbain/rural
Axe 3	+++ % de résidences secondaires -- Nb. personnes / ménage ++ % des log. Meublés ++ Revenu médian --- % de logements vacants -- % des 18-40 ans - Taux de chômage	Opposition communes proches du péri-urbain ou résidentielles / communes présentant des difficultés socio-économiques
Axe 4	+++ Taux d'évolution des ménages +++ Nb. personnes / ménage ++ % de résidences secondaires + Taux de chômage --- % de logements vacants - % des petits logements	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 6 fois pour choisir le maillage qui minimise la variance intra-maille du loyer par m^2.

Comme pour la carte produite en 2020, 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 2 768 mailles, est le plus fin en raison du nombre d'observations plus élevé dans l'échantillon. Il contient 1 741 mailles pour les appartements 1-2 pièces et 1 883 mailles pour les appartements 3 pièces et plus. Celui des maisons comporte 1 956 mailles.

Le maillage des appartements explique 64,3 % de la variance du loyer/m² (62,9 % pour les appartements 1-2 pièces et 77,6 % pour les appartements 3 pièces et plus) contre 51,4 % pour les maisons. À titre de comparaison, la part de variance des loyers/m² expliquée par un maillage communal est légèrement supérieure avec respectivement 64,8 % pour les appartements et 55,3 % 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.

Pour le maillage « appartements », la maille la plus grande compte 67 communes (respectivement 11 EPCI). En moyenne, une maille comprend 13 communes (resp. 3 EPCI).

Pour le maillage « appartements 1-2 pièces », la maille la plus grande compte 128 communes (resp. 17 EPCI). En moyenne, une maille comprend 20 communes (resp. 4 EPCI).

Pour le maillage « appartements 3 pièces et plus », la maille la plus grande compte 127 communes (resp. 11 EPCI). En moyenne, une maille comprend 19 communes (resp. 3 EPCI).

Les mailles du maillage « maisons » sont plus grandes. Elles comptent en moyenne 18 communes et 31 EPCI, et au maximum 96 communes et 6 EPCI.

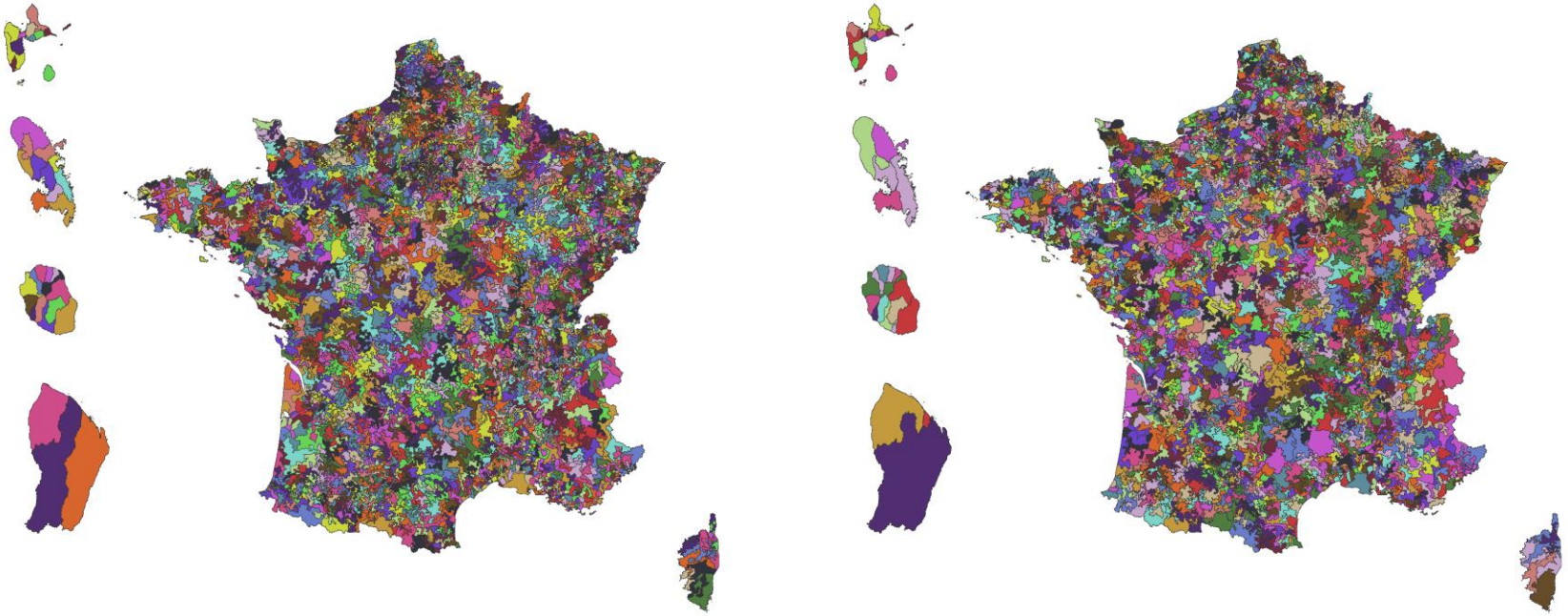
Tableau 9 : Statistiques descriptives sur le nombre d'observations des mailles

Type de biens	Min	Q1	D1	Q2	Moy	Q3	D9	Max	Ecart_type
Appart.	500	575	509	823	2 077	1 592	3 792	123 164	5 238
T1/T2	500	565	507	779	1 878	1 518	3 461	79 268	4 217
T3 et plus	500	535	504	699	1 317	1 146	2 358	44 219	2 245
Maison	500	511	502	577	732	755	1 090	5 587	441

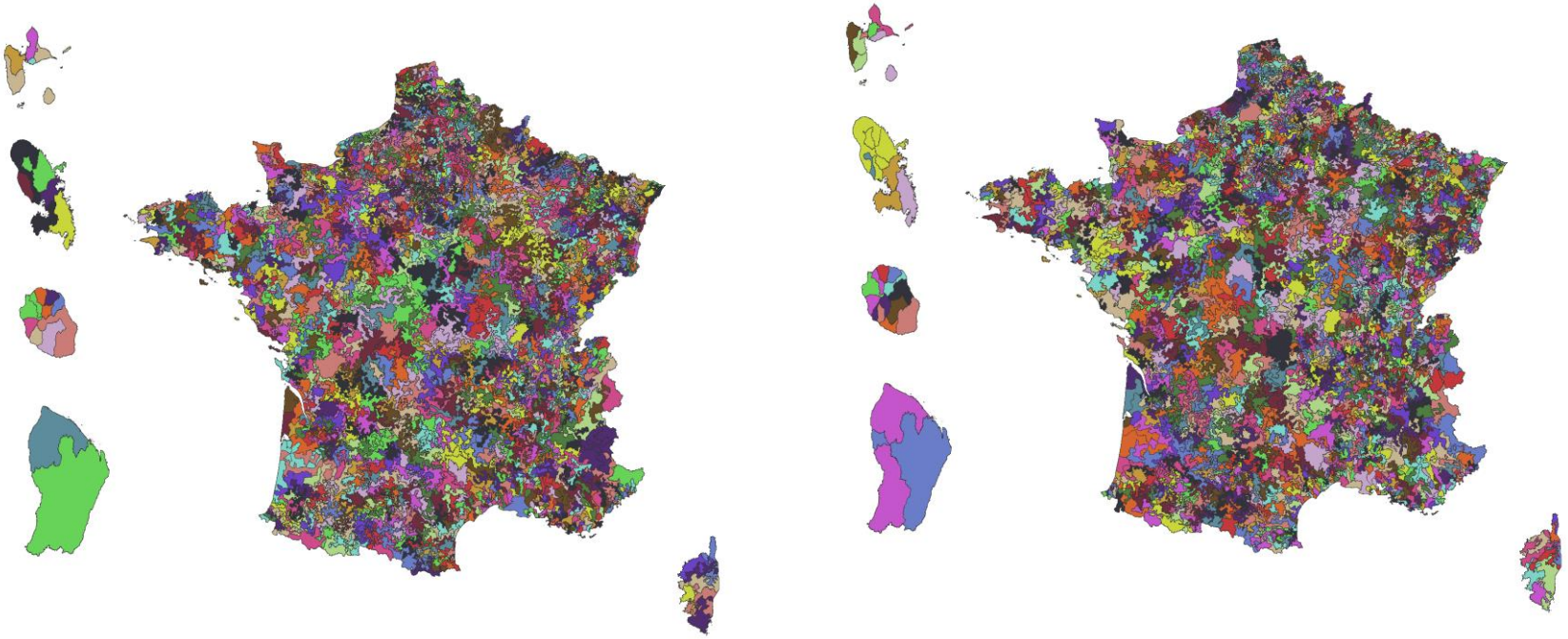
En moyenne, une maille du maillage « appartements » compte 2 077 observations contre seulement 732 pour le maillage « maisons ». Le nombre maximal varie également fortement, avec 123 164 observations pour le maillage « appartements » (à Toulouse) et 5 587 pour le maillage « maisons ».

Les figures 2, 3, 4 et 5 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)



Figures 4 et 5 : Cartographie du maillage pour les appartements 1-2 pièces (à gauche) et 3 pièces et plus (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 pas 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, appartements 1-2 pièces, appartements 3 pièces et plus, 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[5] de la surface en m² ; $SurfMoyPièce$ est la surface moyenne par pièce[6] 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[7]. 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) :

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

[5] 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.
[6] 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.
[7] 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.

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 10. Les R2 moyens s'élèvent à 0,79 pour les

appartements, 0,68 pour les appartements 1-2 pièces, 0,62 pour les appartements 3 pièces et plus et 0,73 pour les maisons. Les R2 minimums sont très faibles pour les appartements 1-2 pièces (0,22) et pour les appartements 3 pièces et plus (0,20) mais plus élevés pour les appartements tous confondus (0,37) et pour les maisons (0,44). Les R2 faibles pour les appartements 1-2 pièces et 3 pièces et plus 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 10 : Statistiques descriptives sur les R2 ajustés obtenus

	min.	**D1**	**Q1**	**Q2**	**moy.**	**Q3**	**D9**	**max.**	**écart-type**
appart.	0.374	0.697	0.744	0.790	0.786	0.834	0.873	0.965	0.071
T1 et T2	0.221	0.553	0.617	0.682	0.676	0.744	0.790	0.920	0.095
T3 et plus	0.202	0.508	0.559	0.619	0.619	0.676	0.733	0.918	0.093
maison	0.443	0.632	0.681	0.729	0.727	0.782	0.819	0.916	0.073

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[8] 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 :

$$indicateur\ de\ loyer_i = \hat{\alpha} + \widehat{\beta_S} f(\overline{Surface}) + \widehat{\beta_{SMP}}(\overline{SurfMoyPIèce}) + \widehat{\beta_T}\ 3\text{ème}Trim + \widehat{\beta_A}\ 2022 + \widehat{\beta_B}\ SeLoger/lbcpro + \widehat{F_i}$$

où $\widehat{\beta_S}$, $\widehat{\beta_{SMP}}$, $\widehat{\beta_T}$, $\widehat{\beta_A}$, $\widehat{\beta_B}$ et $\widehat{F_i}$ sont les coefficients estimés[9].

[8] 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.

[9] 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.

Les caractéristiques du bien de référence retenu pour chaque type de bien sont présentées dans le tableau 11.

Tableau 11 : Caractéristiques retenues pour réaliser les prédictions de loyers

Type de biens	**Appartements**	**Appartements 1-2 pièces**	**Appartements 3 pièces et plus**	**Maisons**
Surf.	52 m²	37 m²	72 m²	92 m²
Surf. moy./pièce	22,2 m²	22,9 m²	21,2 m²	22,3 m²
Trim.	3ème trim.			
Année	2022			
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 52 m^2 et une surface moyenne par pièce de 22,2 m^2 pour les appartements. Les prédictions sont réalisées pour un logement proposé à la location au troisième trimestre de l'année 2022. 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 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.

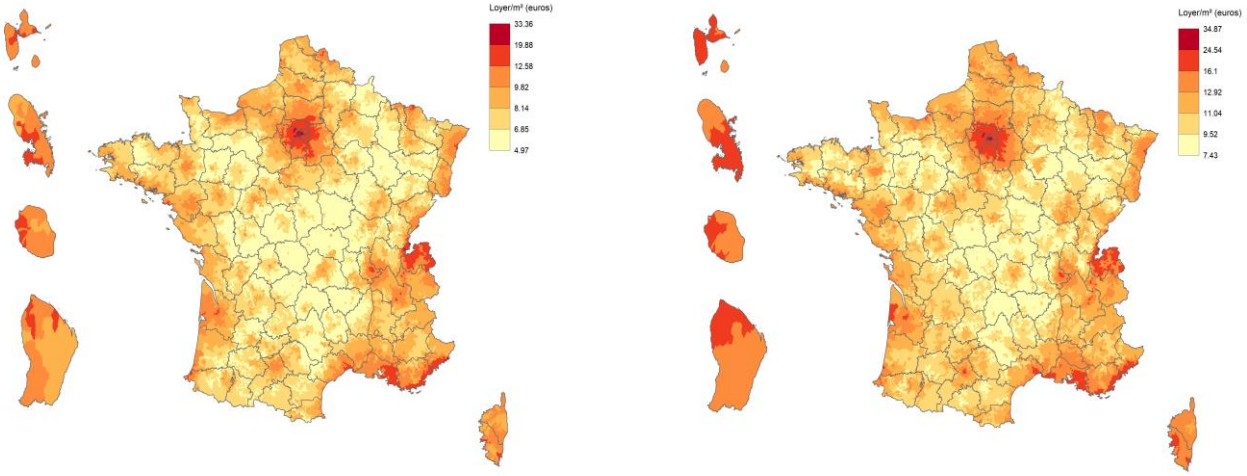
La distribution des indicateurs de loyers obtenus pour ces biens de référence sont présentés dans le tableau 12, et leur répartition géographique est cartographiée par les figures 6, 7, 8 et 9.

Tableau 12 : Distribution des loyers et des loyers/m² prédits

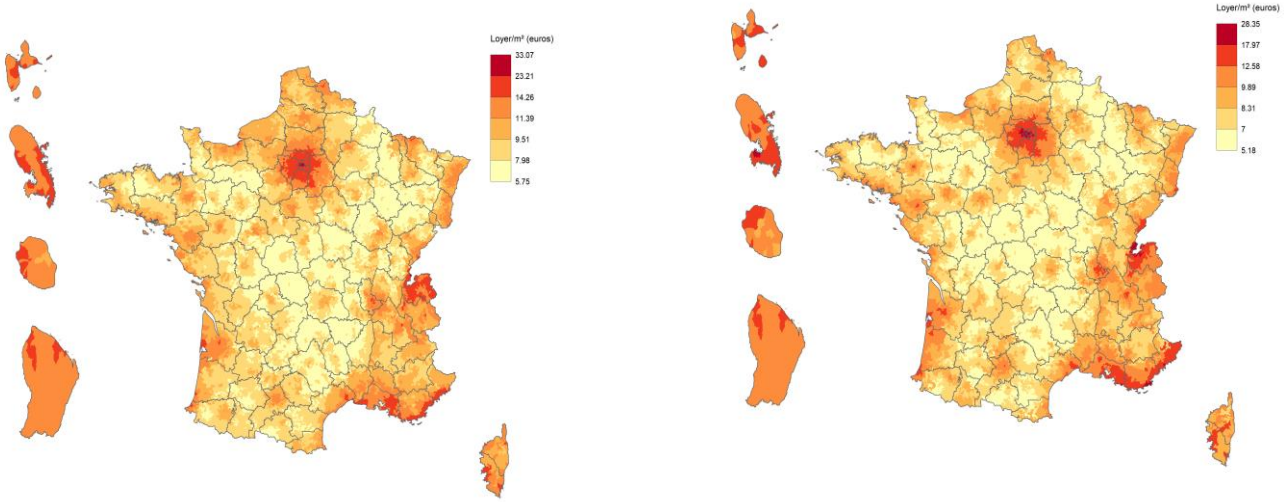
	min	D1	Q1	Q2	moy	Q3	D9	max	écart-type
appart.	299	371	412	466	488	537	626	1720	113
T1 et T2	275	326	359	403	416	451	510	1290	85
T3 et plus	358	441	484	550	583	643	751	2402	146
maison	477	578	631	711	754	822	963	2608	193

	min	D1	Q1	Q2	moy	Q3	D9	max	écart-type
appart.	5.75	7.13	7.93	8.97	9.38	10.32	12.03	33.07	2.18
T1 et T2	7.43	8.81	9.70	10.90	11.24	12.19	13.80	34.87	2.29
T3 et plus	4.97	6.12	6.72	7.64	8.09	8.93	10.43	33.36	2.03
maison	5.18	6.28	6.86	7.73	8.20	8.93	10.47	28.35	2.09

Figures 6 et 7 : Cartographie des loyers/m² prédits pour les appartements (à gauche) et pour les maisons (à droite)



Figures 8 et 9 : Cartographie des loyers/m² prédits pour les appartements 1-2 pièces (à gauche) et pour les appartements 3 pièces et plus (à droite)



E. Analyse des limites et précautions d'usages

1/ Limites

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. La présence des montants de charges dans les données SeLoger pourraient permettre d'étudier la production d'un indicateur hors charges (en procédant à une estimation des charges pour les données leboncoin).

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 communément aux deux bases de données utilisées, alors qu'elles constituent des déterminants importants des loyers. Ceci réduit le pouvoir explicatif du modèle et introduit potentiellement un biais de variables omises si une variable explicative est corrélée avec une variable omise. 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.

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 SeLoger, présentent des caractéristiques similaires dans chacune des mailles de la zone d'étude, ce qui est susceptible d'engendrer un biais de sélection.

2/ Précautions d'usages

Les utilisateurs sont invités à considérer avec prudence les indicateurs de loyer dans les communes où le coefficient de détermination (R2) est inférieur à 0,5, le nombre d'observations dans la commune est inférieur à 30 ou l'intervalle de prédiction est très large.

Il est également à noter que les deux millésimes des cartes (2020 et 2022) ne peuvent pas être comparés, étant donné certaines différences significatives entre ces deux millésimes :

- Dans cette nouvelle version de la carte, les annonces prises en compte sont circonscrites aux locations non meublées ; cette distinction n'était pas encore possible lors de l'élaboration de la première carte ;
- Le maillage d'appartenance des communes a évolué entre les deux millésimes de la carte. L'estimation du loyer n'est donc pas faite « à périmètre constant » dans certains territoires par rapport à la carte diffusée en 2020.
- Les biens types pour lesquels les indicateurs de loyer par m² sont calculés ne sont pas les mêmes dans les deux versions de la carte.

La carte des loyers ne permet donc pas de mesurer des évolutions de loyer au cours du temps. Elle offre une photographie, pour un trimestre, des niveaux de loyer et permet surtout de comparer des territoires entre eux.