Grille des loyers 2021

Données 2017-2018-2020 - proposition d'une nouvelle grille

Pierre Marissal - Hugo Périlleux - Mathieu Strale - ULB-IGEAT

30/08/2021

Contents

I. Données

Pour réaliser cette proposition de grille des loyers, nous nous sommes basés sur les données de 2017, 2018 et 2020. Dans cette section, nous décrivons des prétraitements, nous importons et sélectionnons les données et enfin nous créons les variables synthétiques utiles à la suite des traitements.

1. Pré-traitements

Nous avons réalisé les prétraitements suivants :

- nettoyage des fichiers originaux: ne garder qu'une ligne d'en-tête (noms variables, par questions de l'enquête)
- exportation en format .csv des variables (variables_2017.csv, variables_2018.csv et variables_2020.csv)
- Pour la nouvelle grille, sur base des valeurs 2017, 2018 et 2020, il faut rapporter tout aux prix de 2020. Nous calculons le déflateur à partir de l'indice santé disponible sur https://statbel.fgov.be/fr/open-data/indice-des-prix-la-consommation-et-indice-sante, colonne MS_HLTH_IDX: on fait une approximation de l'inflation par la moyenne de l'année (somme des mois / 12) et on divise la moyenne de l'année en question par la moyenne de l'année de référence pour avoir le facteur de déflation/inflation (2020/2018 = 1,02658775866863; 2020/2017= 1,04422304640471).

2. Importation des données de l'enquête

Nous importons les données de 2017, 2018 et 2020. Il nous semble important de resouligner notre scepticisme vis-à-vis de la qualité des données. En effet, aucune stratification de l'échantillon n'est réalisée, c'est-à-dire qu'il n'y a aucun moyen de vérifier que l'échantillon est représentatif en comparant des données obtenues. Par ailleurs, comme le census, par ailleurs les consignes données aux enquêteurs pour choisir leurs enquêtés sont très peu précises et il semble que trop peu de moyens aient été accordés à l'enquête pour espérer des données de bonne qualité (voir rapport précédent).

3. Sélection des variables, création de variables synthétiques et nettoyage

Nous sélectionnons les variables suivantes:

- loyer (en euro 2020),
- superficie (en m^2),
- type de logements (Studio; Appartement 0 chambre; 1, 2, 3, 4 et plus; Maison 1 chambre; 2, 3, 4 et plus),
- état perçu (de 1 - "Très mauvais état" à 5 - "Très bon état"), nous mettons ensemble dans les états perçus 1, 2 et 3,
- nombre de garages,
- présence d'un convecteur (1 - présence, 0 - absence), Nous utilisons la variable convecteur plutôt que chauffage centrale (variable utilisée dans la grille actuelle) parce que dans le questionnaire la question posée laisse le choix entre 4 possibilités: chauffage central collectif, chauffage central individuel, convecteurs ou de poêles individuels, ou logement passif ou basse énergie. Dès lors, la variable convecteurs permet d'exclure les 3 autres qui offrent un confort supérieur.

Nous construisons les variables synthétiques suivantes:

- variable synthétique d'état (1 - logement sans double vitrage à toutes les fenêtres et construit en 1999 ou avant, 2 - logement avec du double vitrage à toutes les fenêtres et construit en 1999 ou avant, 3 - logement construit en 2000 ou après),

- présence d'une deuxième salle de bain (1 - avec une deuxième salle de bain, 0 - sans deuxième salle de bain),
- absence d'outils de régulation thermique (thermostat et vannes thermostatiques) (1 - absence, 0 - présence) [attention pour l'année 2017, il n'y a aucun logement sans outils de régulation thermique],
- absence d'espace récréatif (balcon, terrasse, cour ou jardin) (1 - absence, 0 - présence),
- présence d'espace de rangement (1 - présence d'un grenier et/ou une cave - 0 absence)

Vu les difficultés opérationnelles à différencier les appartements avec 0 chambre des studios, nous décidons de les regrouper pour la suite de l'étude. Vu la faiblesse de l'échantillon pour les catégories maisons 1 et 2 chambres (voir Table 1), nous décidons aussi de les mettre ensemble.

Table 1: Tableau des effectifs des différents les types de logements

Var1	Freq
Appartement0	301
Appartement1	5240
Appartement2	5535
Appartement3	1375
Appartement4	344
Maison1	53
Maison2	128
Maison3	177
Maison4	254
Studio	635

4. Importation des données sur les quartiers

Comme variable d'environnement, nous utilisons l'indice de difficulté de 2010 qui va de -2 pour les quartiers sans difficulté à 4 pour les quartiers avec beaucoup de difficultés. La figure 1 présente la distribution de la variable pour les quartiers de Bruxelles.

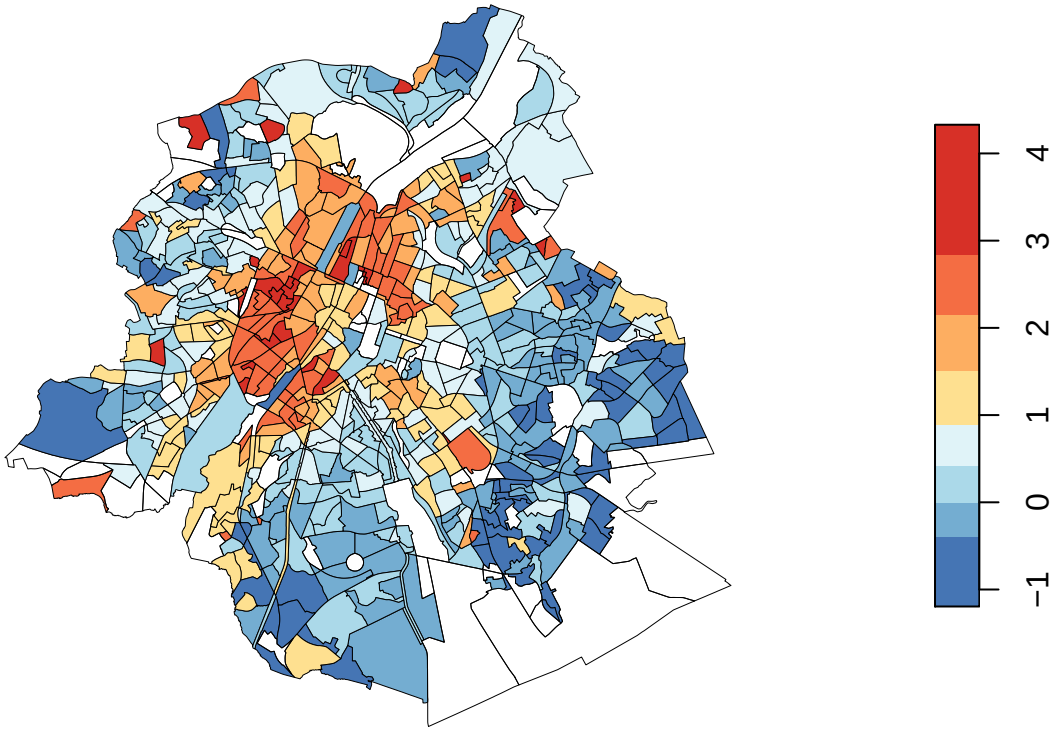

Figure 1: Indice de difficulté 2010

II. Relation générale loyer surface

Dans cette section nous investiguons quelle est la relation entre loyer et surface. Autrement dit, quelle est la forme de la courbe qui lie au mieux les loyers et la surface. Dans un premier temps, nous analysons la relation pour tous les types de logements. On réalise ces analyses en enlevant les valeurs extrêmes définies dans un premier temps comme les observations ayant une surface ou un loyer par surface supérieur ou inférieur respectivement au 99ème décile ou au 1er décile. Dans un second temps, on définira les valeurs extrêmes en calculant les déciles par type de logements.

Dans cette section, nous optons pour une approche empirique. Nous testons les différentes relations possibles entre le loyer par m² et la superficie:

$$\text{relation lineaire : } \frac{loyer}{surface} = \beta_0 + \beta_1 surface$$

$$\text{polynomiale 1 : } \frac{loyer}{surface} = \beta_0 + \beta_1 \frac{1}{surface}$$

$$\text{polynomiale 2 : } \frac{loyer}{surface} = \beta_0 + \beta_1 \frac{1}{surface} + \beta_2 \frac{1}{surface^2}$$

$$\text{polynomiale 3 : } \frac{loyer}{surface} = \beta_0 + \beta_1 \frac{1}{surface} + \beta_2 \frac{1}{surface^2} + \beta_3 \frac{1}{surface^3}$$

$$\text{polynomiale 4 : } \frac{loyer}{surface} = \beta_0 + \beta_1 \frac{1}{surface} + \beta_2 \frac{1}{surface^2} + \beta_3 \frac{1}{surface^3} + \beta_4 \frac{1}{surface^4}$$

Dans un premier temps, nous avons réalisé des régressions suivant la méthode des moindres carré ordinaire (Ordinary Least Squared). Le principe est de faire passer au mieux la droite de régression dans le nuage de points tel que la somme des carrés des résidus soit minimale (formule 1). Les résultats des régressions suivant cette méthode sont des moyennes conditionnelles.

$$\text{(1) OLS: argmin} \sum (y_i - \hat{y_i})^2$$

Dans un second temps, pour respecter l'arrêté selon lequel la grille doit publier des médianes, nous avons réalisé des régressions quantiles. Le principe des régressions quantiles est de faire passer au mieux la droite de régression tel que la somme des valeurs absolues des résidus est minimale (formule 2). Les résultats des régressions suivants cette méthode peuvent être tous les quantiles dont la médiane. Lorsqu'on réalise une régression médiane, la droite passe dans le nuage de point de telle façon que 50% des observations se trouvent en-dessous et 50% au-dessus. Cette méthode de régression a l'avantage de ne pas être influencée par les valeurs extrêmes à l'instar de la médiane.

$$\text{(2) Regression quantile: argmin} \sum |y_i - \hat{y_i}|$$

Vu la proximité des résultats en utilisant les deux méthodes (voir figure 2, figure 4 et table 14) et vu le fait qu'il n'est pas usuel de calculer des R^2 pour les régressions médianes et qu'il nous semble être un bon outils pour estimer la qualité des modèles, nous présentons les prochains les traitements en utilisant la méthode OLS.

1. Graphique général loyer/m² - surface

Grâce à la figure 2, nous voyons que la relation linéaire ne semble pas être adaptée. L'utilisation de l'inverse de la surface semble mieux coller aux données. Étant donné que la relation n'est pas linéaire et que l'utilisation de classes de surface produit des problèmes aux limites des classes (voir rapport précédent), nous décidons de fonctionner par équations pour réaliser la nouvelle "grille des loyers".

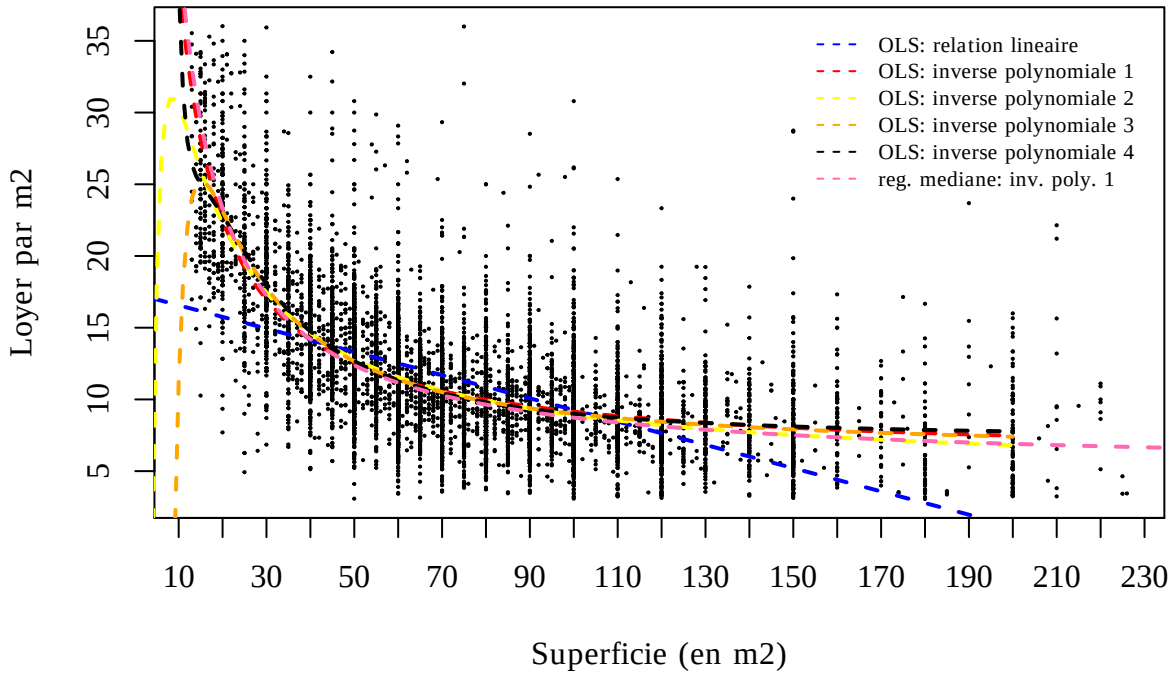

Figure 2: Relation entre loyer et surface

2. Tableau R^2 régression sur l'ensemble des données

Dans cette section nous calculons les R^2 des différents modèles: linéaire, inverse polynomiale 1, 2, 3 et 4.

Table 2: Tableau R^2 pour les différentes relations entre le loyer et la surface

	R^2
linéaire	33.4166
inverse polynomiale 1	55.0809
inverse polynomiale 2	55.9944
inverse polynomiale 3	56.1730
inverse polynomiale 4	56.4200

La table 2 présente les R^2 des différentes régressions. Elle montrent une nette amélioration du R^2 en utilisant l'inverse de la surface plutôt que la relation linéaire. Les gains d'explication avec les polynomiales 2, 3 et 4 sont croissants mais ne permettent au mieux que d'expliquer 1,4% de plus que la polynomiale 1.

3. Problème de multicolinéarité: corrélation entre les types de logements et la surface

Si on décide d'intégrer dans la même régression la surface et le type de logements (studio ou appartement 0 chambres, appartements 1, 2, 3, 4 chambres ou plus, maison 1 ou 2 chambres, maisons 3, 4 chambres ou plus), nous risquons d'avoir un problème de multicolinéarité. En effet, les deux variables semblent a priori très corrélées. Dès lors les estimateurs pourraient être instables.

Table 3: Tableau anova de comparaison des moyennes de surface entre type de biens

Effect	DFn	DFd	F	p	p<.05	ges
type_logement2	7	13044	1397.78	0	*	0.429

Table 4: Tableau test VIF

	GVIF	Df	GVIF^(1/(2*Df))
inv_surface	1.72341	1	1.31279
type_logement2	1.72341	7	1.03964

Au travers du graphique (figure 3) et du test ANOVA (table 3) de comparaison des moyennes de surface entre les différents types de logements, nous voyons qu'il y a bien un lien entre la surface et le type de logements. Le résultat du test VIF (table 4) confirme également le problème de multicolinéarité puisqu'il nous dit que la variance des coefficients est suppérieure de 72,3% à ce qu'on aurait dû observer si les variables n'étaient pas corrélées. Dès lors, pour faire face au problème de multicolinéarité, nous décidons de réaliser les régressions par type de logements et puis de réintégrer les résultats dans une régression générale.

4. Régressions par type de logements

Dans cette section, nous réalisons les régressions par type pour faire face au problème de multicolinéarité soulevé à la section précédente.

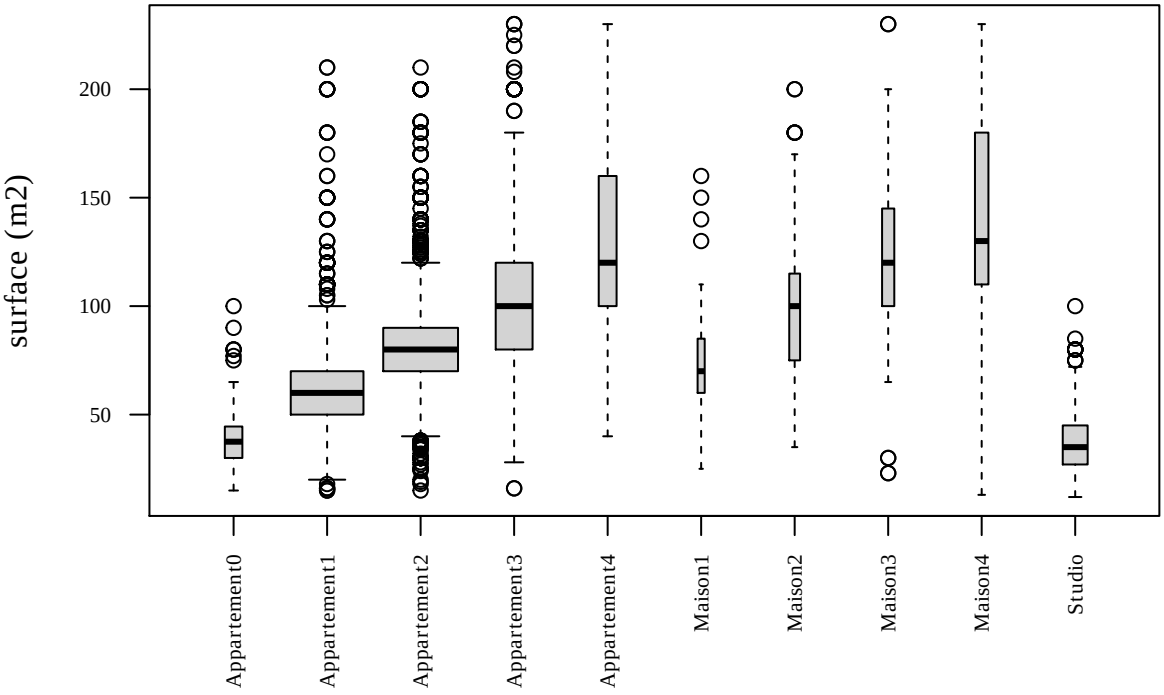

Figure 3: Surface par type de logements

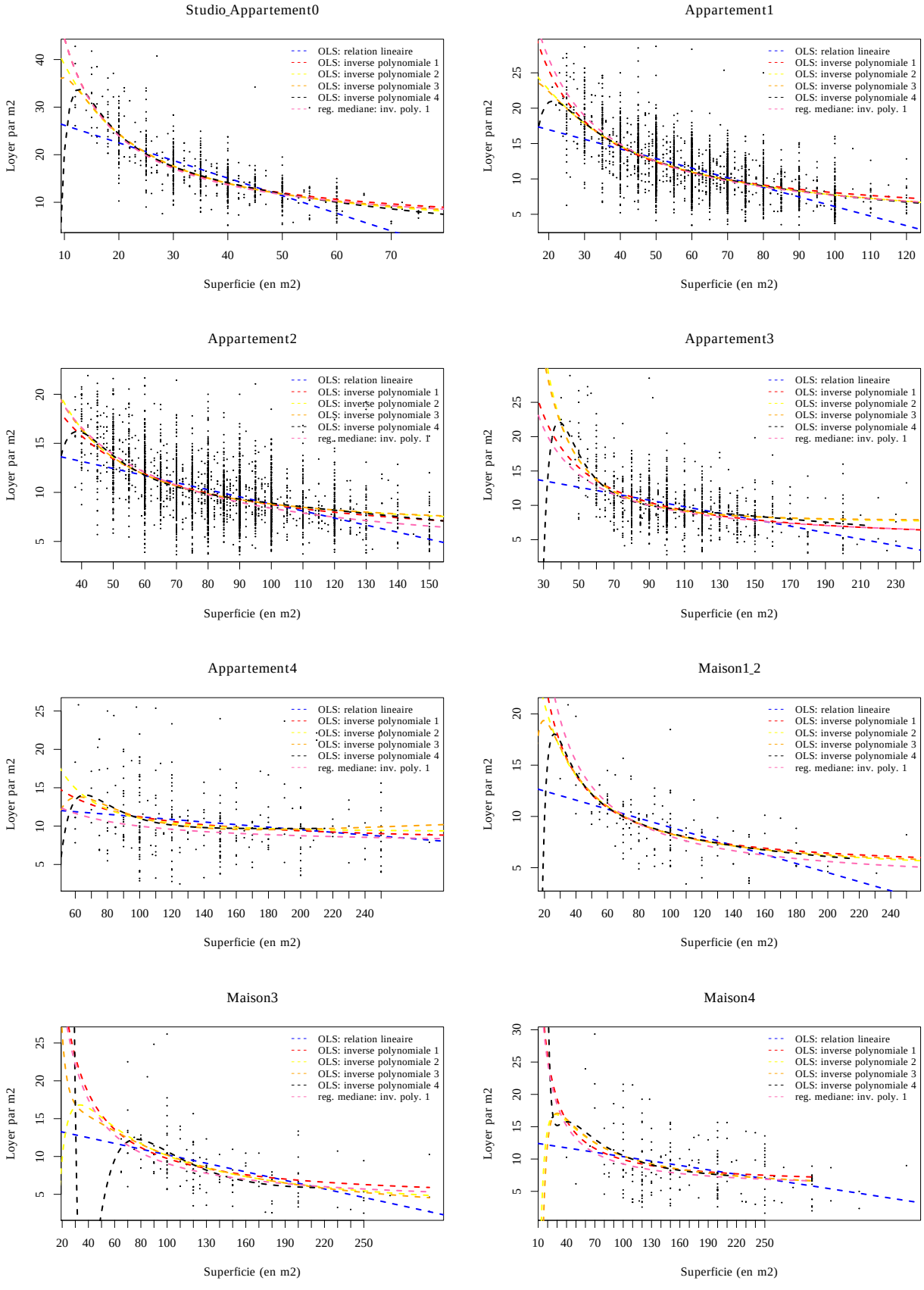

Figure 4: Graphiques des régressions par type de logements

Visuellement (figure 4), il semble que l'utilisation de l'inverse de la surface semble plus adapté que la relation linéaire. Cela dit, il y a peu de différence entre les différentes polynomiales. On observe également que relation varie avec le nombre de chambres: la pente est plus faible pour les logements avec plus de chambres.

Table 5: Tableau des R^2 des différents modèles par type de logements - partie 1

	Studio_Appartement0	Appartement1	Appartement2	Appartement3	Appartement4
lineaire	41.3679	14.0519	17.4916	18.2806	6.83827
poly_1	71.8487	70.7178	67.5113	61.0088	8.71232
poly_2	71.9086	71.2664	68.0441	62.4437	8.71667
poly_3	71.9415	71.2727	69.0732	63.1984	9.02646
poly_4	71.9495	71.3432	69.4302	63.4891	9.21713
lineaire_sans_outliers	53.1300	42.0609	26.3803	19.1457	3.41829
poly_1_sans_outliers	61.4123	46.9087	31.7073	28.9240	5.21914
poly_2_sans_outliers	61.7385	47.2119	31.9498	30.7569	5.55311
poly_3_sans_outliers	61.7638	47.2126	31.9525	30.7705	5.94323
poly_4_sans_outliers	61.9048	47.2207	32.0759	31.5121	6.05313

Table 6: Tableau des R^2 des différents modèles par type de logements - partie 2

	Maison1_2	Maison3	Maison4
lineaire	15.2187	27.2709	9.72097
poly_1	79.0511	42.5238	60.70425
poly_2	79.0681	45.0172	61.89067
poly_3	81.1122	45.0839	64.09838
poly_4	84.4883	45.3413	64.32017
lineaire_sans_outliers	30.5449	24.3647	12.06481
poly_1_sans_outliers	41.3697	25.4110	13.61624
poly_2_sans_outliers	41.5954	28.3327	15.24112
poly_3_sans_outliers	41.6151	28.4305	15.24992
poly_4_sans_outliers	41.7227	30.4930	15.47040

Les R^2 des différentes régressions (table 5 et 6) montrent pour la plupart des types une nette amélioration en utilisant l'inverse de la surface plutôt que la relation linéaire. Étant donné que nous ne pouvons expliquer théoriquement les modèles avec les polynomiales 2, 3 et 4 et que les gains de R^2 sont modérés, nous faisons le choix de la polynomiale 1 avec l'utilisation de l'inverse de la surface pour tous les types.

III. Tests préalables au modèle de régression

L'hypothèse d'homoscédasticité des résidus est essentielle pour la construction de modèles de régression. On dit que l'hypothèse d'homoscédasticité est rencontrée lorsque la variance des résidus est constante le long de la courbe de régression. Cette hypothèse est d'autant plus importante si on souhaite établir des intervalles constants le long de la courbe de régression. Dans ce chapitre, nous analyserons des tests d'homoscédasticité pour des régressions pour chaque type de logements puis nous ferons des tests supplémentaires pour le seul type problématique. Étant donné que les variables d'état et d'environnement permettent d'expliquer une part très faible de la variance (R^2), on teste la constance de la variance des résidus sur les régressions uniquement avec la variable (inverse de la) surface.

1. Test d'homoscédasticité par type

Dans cette section on réalise des test homoscédasticité des résidus pour chacune des régressions par type de logements. Il ressort de la table 7, qu'il n'est pas possible de rejeter l'hypothèse nulle d'hétéroscédasticité pour le type Maison 1 ou 2 chambres. Dès lors, nous poursuivons les analyses pour ce type problématique.

Table 7: Tableau des tests d'homoscédasticité des résidus pour les différentes régression par type de logements

	ChiSquare	Df	p
Studio_Appartement0	80.0793	1	3.5968e-19
Appartement1	439.437	1	1.43564e-97
Appartement2	122.95	1	1.42999e-28
Appartement3	75.3817	1	3.87972e-18
Appartement4	6.68935	1	0.00969903
Maison1_2	0.881882	1	0.347687
Maison3	10.3565	1	0.00129021
Maison4	20.135	1	7.21621e-06

2. Test d'homoscédasticité pour le type problématique

Dans cette section, nous réalisons des test d'homoscédasticité des résidus en prenant de plus en plus de données dans un premier temps, à partir (du bas) de 50 m² jusque 200 m² et dans un second temps à partir (du haut) de 200 m² jusque 50 m².

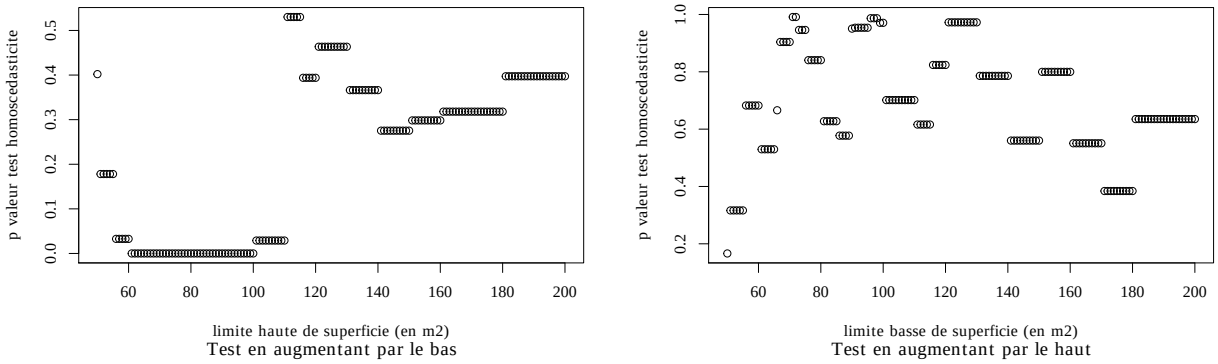

Figure 5: Graphique test homoscédasticité en élargissant le set de données par le bas et par le haut pour le type maison 1 ou 2 chambres

Il ressort du test présenté à la figure 5 que en prenant des maisons 1 ou 2 chambres en dessous de 100 m² les résidus sont homoscédastiques mais pas au-delà (graphique de gauche). Il ressort également qu'il n'est pas pertinent de réaliser une régression pour des maisons 1 ou 2 chambres avec des surfaces supérieures ou égales à 100 m² (graphique de droite). Pour les maisons 1 ou 2 chambres avec une surface supérieure ou égale à 100 m², l'hypothèse d'homoscédasticité n'est pas rencontrée. Dès lors, nous préférons les écarter du modèle général.

IV. Modèle général

Les auteurs de la première grille des loyer et du rapport de 2016 rapportent que les régressions testées avaient des R^2 entre 0,55 et 0,62. Nous n'avons jamais réussi à reproduire leur modèle. Lorsque nous avons tenté de reproduire les régressions avec les variables présentées dans la grille, nous avions obtenu un R^2 de 0,17 (ULB, Rapport grille des loyers, octobre 2020, p8). Dans ce chapitre, nous présentons les résultats de nos régressions.

Dans cette section nous montrons les résultats des régressions OLS qui nous ont guidés dans le choix des variables d'état. Ensuite Nous présentons les résultats avec les régressions quantiles.

1. Régression OLS

A. Modèles par type de logements OLS

Afin de conserver la variable du type de logement et la surface, malgré leur forte corrélation, on décide de réaliser un modèle en deux temps: on réalise d'abord pour chaque type une régression qui vise à expliquer le loyer au m^2 par l'inverse de la surface (étape 1). Les résultats des régressions par type de logements sont présentés dans les tables 8 et 9.

$$\text{(etape 1:) pour chaque type de logements: } \frac{\widehat{loyer}}{surface} = \beta_0 + \beta_1 \frac{1}{surface}$$

Table 8: Tableau régressions par type de logements (étape 1) - partie 1

	Appart. 0 ch.	Appart. 1 ch.	Appart. 2 ch.	Appart. 3 ch.	Appart. 4 ch. et plus
intercept	3.858 ***	3.722 ***	4.086 ***	4.088 ***	7.560 ***
	(0.370)	(0.122)	(0.129)	(0.275)	(0.762)
inverse de la surface	405.987 ***	432.814 ***	466.806 ***	571.109 ***	366.371 ***
	(11.012)	(6.614)	(9.594)	(25.251)	(88.108)
n	856	4849	5101	1259	316
R^2	0.614	0.469	0.317	0.289	0.052

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

B. Modèle général OLS

Ensuite, on prend les loyers par m^2 estimés dans chacune des régressions et on y ajoute la variable d'état et celle d'environnement pour former le modèle général (étape 2). La table 10 présente les résultats des régressions avec différentes combinaisons des variables.

$$\text{(etape 2:) } \frac{\widehat{\widehat{loyer}}}{surface} = \beta_0 + \beta_1 \frac{\widehat{loyer}}{surface} + \beta_2 etat + \beta_3 environement$$

Le modèle (1) avec uniquement les prédictions des modèles par type (voir tables 8 et 9) possède déjà un R^2 de 63,7%. Le modèle (3) avec l'indice synthétique de difficulté et la variable d'état synthétique basée sur la

Table 9: Tableau régressions par type de logements (étape 1)- partie 2

	Maison 1 à 2 ch.	Maison 3 ch.	Maison 4 ch. et plus	NA
intercept	4.528 ***	3.985 ***	5.882 ***	5.962 ***
	(0.481)	(0.764)	(0.581)	(0.813)
inverse de la surface	379.596 ***	575.976 ***	402.742 ***	316.913 ***
	(36.181)	(81.669)	(69.670)	(13.185)
n	158	148	214	531
R^2	0.414	0.254	0.136	0.522

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

présence de double vitrage et l'année de construction possède un R^2 de 65,1% tandis que le modèle (5) avec la variable d'état perçu a un R^2 de 65,5%. La comparaison des modèles 2 et 3 ainsi que des modèles 4 et 5, fait ressortir un gain de R^2 lié à la variable d'environnement de l'ordre de 1,5%. Vu la faible différence de R^2 entre les modèles avec les différentes variables d'état, nous faisons le choix de retenir le modèle (3) avec la variable d'état la plus objective et la plus facilement opérationnalisable. Sur base des faibles gains de R^2 avec les variables d'état et d'environnement vis-à-vis du R^2 obtenu uniquement grâce à l'inverse de la surface par type, nous décidons pour la suite de n'utiliser que la surface comme référence pour réaliser les intervalles de confiance.

2. Régression quantile

A. Modèles par type de logements

Comme avec les régressions OLS, on réalise d'abord pour chaque type une régression qui vise à expliquer le loyer au m² par l'inverse de la surface (étape 1). Les résultats des régressions médiane par type de logements sont présentés dans les tables 11 et 12.

$$\text{(etape 1:) pour chaque type de logements: } \frac{\widehat{loyer}}{surface} = \beta_0 + \beta_1 \frac{1}{surface}$$

Table 10: Modèles généraux pour expliquer le loyer par m² avec les prédictions des régressions par type de logements, une variable d'état et une variable d'environement (étape 2)

	(1)	(2)	(3)	(4)	(5)
intercept	0.000	-0.304 **	0.177	-0.414 ***	0.138
	(0.079)	(0.100)	(0.100)	(0.088)	(0.090)
prédiction modèle par type	1.000 ***	1.005 ***	1.022 ***	1.008 ***	1.022 ***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)
dbl. vitr. et constr. < 2000		0.268 ***	0.241 ***		
		(0.065)	(0.064)		
constr. >= 2000		1.247 ***	1.113 ***		
		(0.131)	(0.129)		
état perçu '4. Bien'				0.442 ***	0.305 ***
				(0.057)	(0.057)
état perçu '5. Très bien'				0.849 ***	0.689 ***
				(0.077)	(0.076)
indice synth. de difficulté			-0.655 ***		-0.640 ***
			(0.029)		(0.028)
n	13432	12261	12261	13315	13315
R^2	0.637	0.636	0.651	0.641	0.655

*** p < 0.001; ** p < 0.01; * p < 0.05.

B. Modèle général médiane

Ensuite, on prend les loyers par m² estimés dans chacune des régressions et on y ajoute la variable d'état et celle d'environnement pour former le modèle général (étape 2). La table 13 présente les résultats de la régression médiane pour la seconde étape du modèle.

$$\text{(etape 2:)} \ \frac{\widehat{\widehat{loyer}}}{surface} = \beta_0 + \beta_1 \frac{\widehat{loyer}}{surface} + \beta_2 etat + \beta_3 environement$$

Il n'est pas usuel de calculer des R^2 pour les régressions quantiles parce que ce n'est pas le critère utilisé. Néanmoins, afin d'avoir un point de comparaison avec les autres régressions nous avons calculé le R^2 de ce modèle général utilisant les régressions médiane : 0.76 .

Table 11: Tableau régressions médiane (étape 1)- partie 1

	Appart. 0 ch.	Appart. 1 ch.	Appart. 2 ch.	Appart. 3 ch.	Appart. 4 ch. et plus
intercept	3.40 ***	2.83 ***	2.91 ***	4.40 ***	7.50 ***
	(0.41)	(0.16)	(0.15)	(0.38)	(0.78)
inverse de la surface	410.94 ***	482.69 ***	548.16 ***	505.96 ***	250.00 *
	(13.69)	(9.87)	(13.25)	(38.38)	(97.91)
N	887	4980	5248	1295	331
tau	0.50	0.50	0.50	0.50	0.50
R1	0.44	0.33	0.24	0.20	0.02
AIC	4867.83	23832.54	23924.15	6631.56	1968.21
BIC	4877.40	23845.57	23937.28	6641.89	1975.81

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Table 14: Tableau de corrélation des estimations des modèles testés

	OLS inv. poly. 1	Rég. médiane inv. poly.1	Rég. médiane lin
OLS inv. poly. 1	1.0000	0.9961	0.9326
Rég. médiane inv. poly.1	0.9961	1.0000	0.7094
Rég. médiane lin	0.9326	0.7094	1.0000

Table 12: Tableau régressions médiane (étape 2)- partie 2

	Maison 1 à 2 ch.	Maison 3 ch.	Maison 4 ch. et plus	NA
intercept	3.17 ***	3.45 ***	5.30 ***	4.44 ***
	(0.76)	(0.48)	(1.01)	(0.62)
inverse de la surface	487.90 ***	562.19 ***	393.88 **	328.93 ***
	(57.32)	(43.01)	(142.87)	(13.24)
N	163	155	223	547
tau	0.50	0.50	0.50	0.50
R1	0.33	0.23	0.13	0.40
AIC	769.17	816.05	1322.65	3726.17
BIC	775.36	822.13	1329.46	3734.78

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

V. Variables additionnelles

1. Régression sur les résidus absolus

Enfin, nous avons réalisé une régression OLS présentée dans la table 15 qui vise à expliquer les résidus absolus (en €) du modèle général par les variables "additionnelles". Il ressort que toutes les variables testées sont significatives mais le R^2 de l'ensemble de la régression est de seulement 2,9%. Seule la variable deuxième salle de bain possède un coefficient important (+118€). Le R^2 associé à la régression des résidus absolus par la variable de présence d'une seconde salle de bain est de 2,1%. Dans notre modèle, les autres variables n'influencent que très peu le loyer.

2. Analyse en composante principale

Il est possible que les résultats présentés plus haut (table 15) soient influencés par le fait que la variable présence d'une deuxième salle de bain masque les autres variables (par le fait qu'elles soient corrélées). Pour tester ce problème potentiel, suivant la suggestion de l'IBSA dans le comité d'accompagnement, nous avons réalisé une régression sur les composantes issues de l'ACP qui sont par définition indépendantes entres elles. Les tables 16 et 17 présentent les corrélations entre chacune des variables. Les corrélations sont très faibles. On voit néanmoins des valeurs non négligeables entre deux groupes de variables: d'une part la présence d'une deuxième salle de bain, l'absence d'espace récréatif et la présence de garage, d'autre part la présence d'un convecteur et l'absence de régulation thermique. La figure 6 présente le pourcentage de variance expliqué par chacune des composantes (dimensions) issues de l'ACP que nous avons réalisée sans centrer-réduire les variables. Nous avons fait ce choix pour donner davantage de poids aux variables qui ont le plus de variance. Nous avons utilisé ces composantes pour tenter d'expliquer les résidus absolus au travers d'une régression OLS (table 20). Nous avons ensuite réparti les coefficients au travers d'une somme pondérée sur base des contributions de chaque variable à chacune des dimensions (table 19) et du signe de la corrélations entre chacune des variables et dimensions (table 18). Les nouveaux coefficients (table 21) obtenus sur base de cette procédure montrent quelques changements vis-à-vis de la première régression (table 15): le coefficient de la présence d'une seconde salle de main passe de 118,5 à 88,5, pour le nombre de garages de 17,6 à 40,1

Table 13: Tableau régressions médianes

	Model 1
intercept	0.18
	(0.11)
prédiction modèle par type	1.02 ***
	(0.01)
dbl. vitr. et constr. < 2000	0.25 ***
	(0.05)
constr. >= 2000	1.04 ***
	(0.10)
indice synth. de difficulté	-0.65 ***
	(0.02)
N	12599
tau	0.50
R1	0.40
AIC	61656.25
BIC	61693.45

*** p < 0.001; ** p < 0.01; * p < 0.05.

et l'absence d'espace de récréatif de -13,6 à -15,7. Ceci confirme que la variable présence d'une seconde salle de bain masquait bel et bien les autres variables. Cela dit, même après ces traitements, les coefficients restent mesurés par rapport aux loyers estimés. Ceci nous amène à suggérer de ne pas utiliser ces variables additionnelles pour ne pas donner l'illusion d'une (fausse) précision du modèle.

La seule variable pour laquelle la question reste posée est la présence d'une seconde salle de bain. En effet, ni coefficient (+118€) , ni le R^2 associé (1,8%) ne sont négligeable. Nous suggérons néanmoins de ne pas retenir cette variable parce que nous n'avons pas épuisé la question des variables additionnelles et qu'il nous semble que ce serait mal venu de donner l'impression d'avoir un modèle sophistiquée et faussement précis.

Table 16: Tableau de correlation des variables additionnelles - partie 1

	deuxieme_salle_bain	garage	convecteur
deuxieme_salle_bain	1.00	0.20	0.00
garage	0.20	1.00	0.00
convecteur	0.00	0.00	1.00
abs_espace_recreatif	-0.13	-0.19	-0.03
abs_regul_thermique	-0.04	-0.06	0.19
rangement	0.11	0.17	0.02

Table 15: Régression sur les résidus absolus avec les variables additionnelles

	(1)	(2)
intercept	-10.287 ***	0.854
	(2.096)	(4.415)
deuxième salle de bain	118.219 ***	118.464 ***
	(7.424)	(8.245)
nombre de garages		17.355 ***
		(5.067)
convecteur		-20.759 *
		(8.757)
absence d'espace récréatif		-13.641 **
		(4.827)
absence de régulation thermique		-24.733 **
		(8.644)
présence espace de rangement		-6.944
		(4.678)
n	11689	10019
R^2	0.021	0.028

*** p < 0.001; ** p < 0.01; * p < 0.05.

Table 17: Tableau de correlation des variables additionnelles - partie 2

	abs_espace_recreatif	abs_regul_thermique	rangement
deuxieme_salle_bain	-0.13	-0.04	0.11
garage	-0.19	-0.06	0.17
convecteur	-0.03	0.19	0.02
abs_espace_recreatif	1.00	0.01	-0.25
abs_regul_thermique	0.01	1.00	-0.02
rangement	-0.25	-0.02	1.00

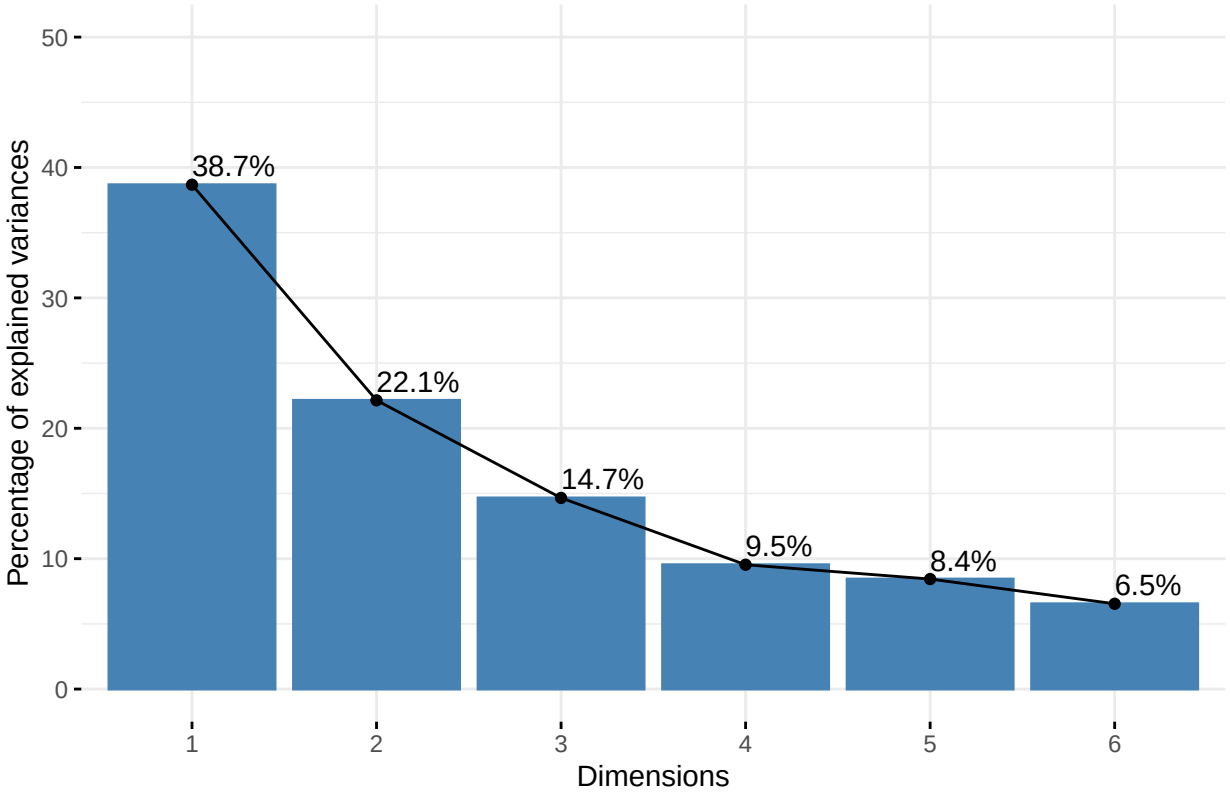

Figure 6: Pourcentage de la variance expliqué par les différentes composantes

Table 18: Tableau de correlation entre les dimmensions et les variables

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6
deuxieme_salle_bain	0.23	-0.05	0.35	-0.07	0.90	-0.04
garage	0.36	-0.08	0.90	0.11	-0.22	-0.01
convecteur	0.04	-0.01	-0.06	0.60	0.10	0.79
abs_espace_recreatif	-0.77	0.62	0.16	0.01	0.02	0.01
abs_regul_thermique	-0.04	0.00	-0.14	0.87	0.07	-0.46
rangement	0.80	0.60	-0.09	0.00	-0.01	0.00

Table 19: Tableau des contributions des variables aux différentes composantes (en pourcent)

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6
deuxieme_salle_bain	1.26	0.09	7.84	0.45	90.15	0.22
garage	5.17	0.41	84.03	1.96	8.41	0.02
convecteur	0.03	0.01	0.19	27.92	0.80	71.05
abs_espace_recreatif	44.16	50.66	4.98	0.05	0.14	0.02
abs_regul_thermique	0.03	0.00	1.22	69.62	0.45	28.68
rangement	49.35	48.84	1.74	0.00	0.06	0.01

Table 21: Tableau des coefficients une fois les coefficients redistribués selon les poids des contributions aux différentes dimmensions (en €)

	coefficients
deuxieme_salle_bain	88.547325
nbre_garage	40.109301
convecteur	-18.679544
abs_espace_recreatif	-15.763757
abs_regul_thermique	-16.867348
rangement	0.707585

VI. Création des intervalles

Dans cette section, nous construisons les intervalles de loyer/m² par type tels qu'ils contiennent 60%, 70%, 80% ou 90% de l'échantillon autour des valeurs prédites par les régressions pour chaque type de logements. Les tables 22 et 23, présentent les bornes inférieures et supérieures aux déciles 0.05, 0.1, 0.15, 0.2, 0.8, 0.85, 0.90, 0.95.

Table 20: Régression sur les résidus absolus avec les différentes dimmensions

	(1)
(Intercept)	-0.542
	(1.943)
Dim.1	21.478 ***
	(3.583)
Dim.2	-18.144 ***
	(4.709)
Dim.3	56.312 ***
	(5.824)
Dim.4	-30.151 ***
	(6.978)
Dim.5	92.825 ***
	(7.673)
Dim.6	-15.346
	(8.550)
n	12261
R^2	0.025

*** p < 0.001; ** p < 0.01; * p < 0.05.

Table 22: Prix au m² à retirer aux valeurs estimées par les équations par type de logements pour obtenir les bornes inférieures de l'intervalle à l'intérieur duquel sont comprises 60, 70, 80, 90 pourcent des données - partie 1

	borne_inf_0.2	borne_inf_0.15	borne_inf_0.1	borne_inf_0.05
Studio_Appartement0	-2.70	-3.24	-4.12	-5.77
Appartement1	-1.80	-2.24	-2.75	-3.74
Appartement2	-1.61	-1.96	-2.44	-3.28
Appartement3	-2.15	-2.51	-2.95	-3.83
Appartement4	-3.30	-3.60	-4.63	-5.81
Maison1_2_<100	-1.94	-2.54	-2.68	-3.54
Maison1_2_>=100	-1.99	-2.08	-2.49	-3.05
Maison3	-2.33	-2.48	-3.17	-4.75
Maison4	-3.22	-3.81	-5.01	-5.80

Table 23: Prix au m² à ajouter aux valeurs estimées par les équations par type de logements pour obtenir les bornes supérieures de l'intervalle à l'intérieur duquel sont comprises 60, 70, 80, 90 pourcent des données - partie 2

	borne_sup_0.8	borne_sup_0.85	borne_sup_0.9	borne_sup_0.95
Studio_Appartement0	2.30	2.93	3.97	6.08
Appartement1	1.79	2.27	2.93	4.09
Appartement2	1.47	1.95	2.60	3.78
Appartement3	1.99	2.61	3.29	4.96
Appartement4	3.19	4.69	6.16	8.21
Maison1_2_<100	1.63	1.99	2.64	3.57
Maison1_2_>=100	2.13	2.41	3.61	4.15
Maison3	1.54	2.71	3.92	6.15
Maison4	3.23	4.61	5.92	7.43

VII. Résumé des équations

Dans cette section, nous présentons les différentes équations par type de logements obtenues avec les modèles de régression médiane, l'inverse de la surface, l'état (présence de double vitrage et année de construction avant ou après 2000) et la variable de localisation (indice de difficulté par quartier). Pour rappel, la variable synthétique d'état vaut: 1 si le logement est sans double vitrage à toutes les fenêtres et construit en 1999 ou avant, 2 si le logement est avec du double vitrage à toutes les fenêtres et construit en 1999 ou avant, 3 si le logement est construit en 2000 ou après.

- Studio-Appartement 0 chambre

loyer = (0.1758082 + 1.0207648* (3.4017754+410.93786 * 1/surface)+ 0.2490667 (si état=2) + 1.042853 (si état=3) -0.6455585 * indice synth. de difficulté) * surface

- Appartement 1 chambre

loyer = (0.1758082 + 1.0207648* (2.8301143+482.6853574 * 1/surface)+ 0.2490667 (si état=2) + 1.042853 (si état=3) -0.6455585 * indice synth. de difficulté) * surface

- Appartement 2 chambres

loyer = (0.1758082 + 1.0207648* (2.9097312+548.1594066 * 1/surface)+ 0.2490667 (si état=2) + 1.042853 (si état=3) -0.6455585 * indice synth. de difficulté) * surface

- Appartement 3 chambres

loyer = (0.1758082 + 1.0207648* (4.3996618+505.9611096 * 1/surface)+ 0.2490667 (si état=2) + 1.042853 (si état=3) -0.6455585 * indice synth. de difficulté) * surface

- Appartement 4 chambres et plus

loyer = (0.1758082 + 1.0207648* (7.5+250 * 1/surface)+ 0.2490667 (si état=2) + 1.042853 (si état=3) -0.6455585 * indice synth. de difficulté) * surface

- Maison 1 ou 2 chambres

loyer = (0.1758082 + 1.0207648* (3.1738354+487.9031965 * 1/surface)+ 0.2490667 (si état=2) + 1.042853 (si état=3) -0.6455585 * indice synth. de difficulté) * surface

- Maison 3 chambres

loyer = (0.1758082 + 1.0207648* (3.4543796+562.1917377 * 1/surface)+ 0.2490667 (si état=2) + 1.042853 (si état=3) -0.6455585 * indice synth. de difficulté) * surface

- Maison 4 chambres ou plus

loyer = (0.1758082 + 1.0207648* (5.300474+393.8815801 * 1/surface)+ 0.2490667 (si état=2) + 1.042853 (si état=3) -0.6455585 * indice synth. de difficulté) * surface

VIII. Loyers “abusifs”

Les loyers auraient été définis comme “abusifs” par le haut (abusivement trop élevés) s'ils dépassent de 20% le loyer estimé. Avec cette définition, 17.74% de notre échantillon respecte le critère de loyer abusif par le haut. Si on définit les loyers “abusifs” par le bas (abusivement trop bas) comme ceux étant 30% en dessous du loyer estimé, 7.46% de l'échantillon respecte le critère de loyer abusif par le bas.

IX. Valeurs exemples sur base des équations

Afin de permettre une meilleure lisibilité du modèle proposé, nous présentons dans cette section une grille avec des valeurs exemples sur base des équations présentées au chapitre précédent. Par exemple un appartement 2 chambres de 100m² dans un quartier avec peu de difficultés (q=-1), ayant été construit après 2000 (état=3) a un loyer prédit de 1089€. Les bornes comprenant 80% de l'échantillon, c'est-à-dire la borne inférieure à 10 % et la borne supérieur à 90% sont estimées respectivement à 1089€-244€= 845€ et 1089€+259€=1348€.

Table 24: Grille avec valeurs d'exemple - partie 1

type_exemple	surface_exemple	borne_inf_0.2	borne_inf_0.15	borne_inf_0.1	borne_inf_0.05
Appartement1	40	72	90	110	150
Appartement1	60	108	134	165	224
Appartement1	80	144	179	220	299
Appartement2	60	97	118	146	197
Appartement2	80	129	157	195	262
Appartement2	100	161	196	244	328
Appartement2	120	193	235	293	394
Appartement3	80	172	201	236	306
Appartement3	100	215	251	295	383
Appartement3	120	258	301	354	460
Appartement3	140	301	351	413	536
Appartement4	100	330	360	463	581
Appartement4	150	495	540	694	871
Maison3	100	233	248	317	475
Maison3	150	349	372	475	712
Maison3	200	466	496	634	950
Maison4	100	322	381	501	580
Maison4	150	483	571	751	870
Maison4	200	644	762	1002	1160
Studio_Appartement0	60	162	194	247	346
Studio_Appartement0	20	54	65	82	115
Studio_Appartement0	40	108	130	165	231

Table 25: Grille avec valeurs d'exemple - partie 2

type_exemple	surface_exemple	borne_sup_0.8	borne_sup_0.85	borne_sup_0.9	borne_sup_0.95
Appartement1	40	72	91	117	164
Appartement1	60	107	136	176	245
Appartement1	80	143	182	234	327
Appartement2	60	88	117	156	227
Appartement2	80	118	156	208	302
Appartement2	100	147	195	260	378
Appartement2	120	176	234	312	454
Appartement3	80	159	209	263	397
Appartement3	100	199	261	329	496
Appartement3	120	239	313	395	595
Appartement3	140	279	365	461	694
Appartement4	100	319	469	616	821
Appartement4	150	478	703	924	1231
Maison3	100	154	271	392	615
Maison3	150	231	406	588	922
Maison3	200	308	542	784	1230
Maison4	100	323	461	592	743
Maison4	150	484	691	888	1114
Maison4	200	646	922	1184	1486
Studio_Appartement0	60	138	176	238	365
Studio_Appartement0	20	46	59	79	122
Studio_Appartement0	40	92	117	159	243

Table 26: Grille avec valeurs d'exemple - partie 3

type_exemple	surface_exemple	q:-1 etat:1	q:-1 etat:2	q:-1 etat:3	q:0 etat:1	q:0 etat:2	q:0 etat:3
Appartement1	40	641	651	683	615	625	657
Appartement1	60	715	730	778	677	692	739
Appartement1	80	790	809	873	738	758	821
Appartement2	60	787	802	850	748	763	811
Appartement2	80	863	883	946	811	831	895
Appartement2	100	939	964	1043	874	899	978
Appartement2	120	1015	1044	1140	937	967	1062
Appartement3	80	941	961	1025	890	910	973
Appartement3	100	1048	1073	1152	983	1008	1087
Appartement3	120	1154	1184	1279	1076	1106	1202
Appartement3	140	1260	1295	1406	1170	1205	1316
Appartement4	100	1103	1128	1207	1038	1063	1143
Appartement4	150	1527	1564	1683	1430	1467	1586
Maison3	100	1009	1034	1113	944	969	1048
Maison3	150	1226	1263	1382	1129	1167	1286
Maison3	200	1443	1493	1652	1314	1364	1523
Maison4	100	1025	1050	1130	961	986	1065
Maison4	150	1337	1374	1493	1240	1277	1396
Maison4	200	1648	1698	1857	1519	1569	1728
Studio_Appartement0	60	677	692	740	638	653	701
Studio_Appartement0	20	505	510	526	492	497	513
Studio_Appartement0	40	591	601	633	565	575	607

Table 27: Grille avec valeurs d'exemple - partie 4

type_exemple	surface_exemple	q:1 etat:1	q:1 etat:2	q:1 etat:3	q:2 etat:1	q:2 etat:2	q:2 etat:3
Appartement1	40	589	599	631	564	574	605
Appartement1	60	638	653	700	599	614	662
Appartement1	80	686	706	770	635	655	718
Appartement2	60	710	725	772	671	686	733
Appartement2	80	760	779	843	708	728	791
Appartement2	100	810	834	914	745	770	849
Appartement2	120	860	889	985	782	812	907
Appartement3	80	838	858	922	787	806	870
Appartement3	100	919	944	1023	854	879	958
Appartement3	120	999	1029	1124	922	951	1047
Appartement3	140	1079	1114	1225	989	1024	1135
Appartement4	100	974	999	1078	909	934	1014
Appartement4	150	1333	1370	1490	1236	1274	1393
Maison3	100	880	904	984	815	840	919
Maison3	150	1032	1070	1189	935	973	1092
Maison3	200	1185	1235	1394	1056	1106	1265
Maison4	100	896	921	1000	832	856	936
Maison4	150	1143	1181	1300	1046	1084	1203
Maison4	200	1390	1440	1599	1261	1311	1470
Studio_Appartement0	60	600	615	662	561	576	623
Studio_Appartement0	20	480	485	500	467	472	487
Studio_Appartement0	40	540	550	581	514	524	555

Table 28: Grille avec valeurs d'exemple - partie 5

type_exemple	surface_exemple	q:3 etat:1	q:3 etat:2	q:3 etat:3
Appartement1	40	538	548	580
Appartement1	60	560	575	623
Appartement1	80	583	603	666
Appartement2	60	632	647	695
Appartement2	80	656	676	740
Appartement2	100	680	705	785
Appartement2	120	705	735	830
Appartement3	80	735	755	818
Appartement3	100	789	814	894
Appartement3	120	844	874	969
Appartement3	140	899	934	1045
Appartement4	100	845	870	949
Appartement4	150	1139	1177	1296
Maison3	100	750	775	855
Maison3	150	839	876	995
Maison3	200	927	977	1135
Maison4	100	767	792	871
Maison4	150	950	987	1106
Maison4	200	1132	1182	1341
Studio_Appartement0	60	522	537	585
Studio_Appartement0	20	454	459	475
Studio_Appartement0	40	488	498	530

X. Comparaison grille actuelle

Afin de comparer les estimations obtenues avec notre modèle et la grille actuelle, nous avons pris l'exemple des appartements une chambre (voir figure 7). Dans la grille actuelle, pour un logement en mauvais état (état=0) et dans un quartier pauvre (Q1), le loyer est de 8,8€/m² (ligne rouge pointillée). Pourtant, on voit sur le graphique que les petits logements ont un loyer/m² bien supérieur à cette valeur. Pour un logement de 35m² en mauvais état dans un quartier pauvre, nos estimations donne un loyer de l'ordre de 15€/m² (ligne rouge pleine), soit quasi le double de ce que prédit la grille actuelle. On passe d'une estimation de environ 310€ avec la grille actuelle à environ 525€ avec notre modèle, ce qui semble bien mieux correspondre aux données. Ce problème est également présent pour les logements en bon état dans les quartiers riches (voir lignes bleues).

Pour résumer, la grille actuelle sous-estime largement les loyers des petits logements. Ceci est dû au fait qu'elle ne tient compte de la surface que au travers du type de bien (appartement ou maison) et du nombre de chambres. Notre modèle résout ce problème en tenant compte explicitement de la surface dans le calcul du loyer/m².

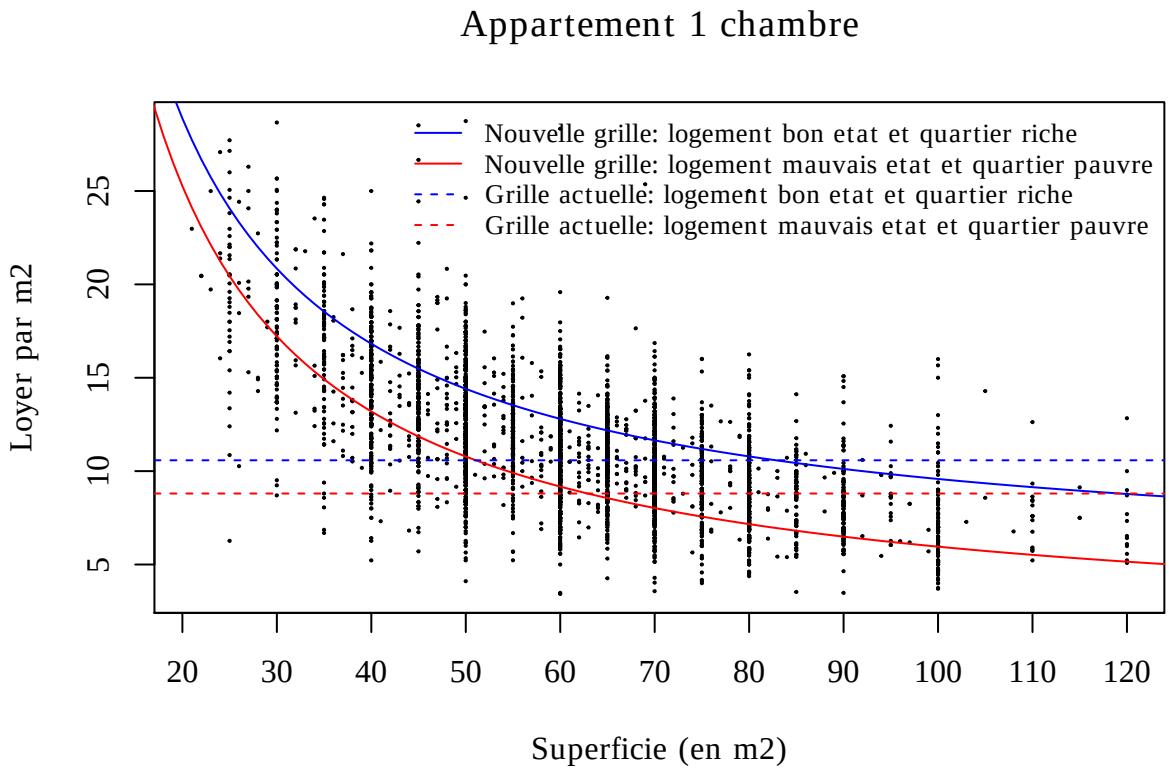

Figure 7: Graphique grille actuelle vs nouvelles estimations