Le bonheur est dans le prix

Estimation de la valeur du patrimoine immobilier des ménages à partir de données exhaustives

Mathias ANDRÉ Olivier MESLIN

Insee, Département des études économiques (DESE)

Journées de Méthodologie Statistique – 31 mars 2022

Résultats provisoires – ne pas diffuser

Plan

Un projet innovant sur le patrimoine immobilier des ménages

Motivation : vers une base exhaustive sur le patrimoine immobilier des ménages

Les inégalités immobilières ont augmenté sur longue période :

- En 40 ans, le niveau de vie des locataires a décroché : leur revenu par UC a augmenté moins vite que les propriétaires, les loyers ont progressé plus fortement que leurs revenus ;
- Le taux de propriétaires parmi les 10 % les plus modestes est passé de plus de 45 % en 1973 à 20 % en 2013.

Notre projet est à la rencontre de deux évolutions :

- Disponibilité croissante de données administratives : développement de sources exhaustives sur les ménages dans la statistique publique (fichiers Fidéli et Filosofi).
- Fort intérêt pour les inégalités de patrimoine : importance des extrémités de la distribution et disparités géographiques.

Un projet aux contributions multiples

Motivation : vers une base exhaustive sur le patrimoine immobilier des ménages

L'exploitation de sources administratives exhaustives permet d'apporter de nouveaux résultats :

1. Étude détaillée de la concentration de la propriété immobilière sur le champ des logements, et lien avec la distribution des revenus ;
2. Analyse redistributive de la taxe foncière sur les propriétés bâties sur l'ensemble du patrimoine d'habitation (résidences principales et secondaires, biens mis en location, biens possédés en SCI) ;
3. Modélisation des prix immobiliers à partir des transactions pour estimer la valeur de marché des logements et le patrimoine immobilier des ménages.

- Disparités territoriales et hétérogénéité en niveau de vie étudiées simultanément.

Deux études déjà publiées

- André, M. et Meslin, O., Et pour quelques appartements de plus : Étude de la propriété immobilière des ménages et du profil redistributif de la taxe foncière, Insee, *Document de travail* n ° 2021-004, novembre 2021
- André, M., Arnold, C., et Meslin, O., 24 % des ménages détiennent 68% des logements possédés par des particuliers, Insee références *France, portrait social*, novembre 2021

Des apports pour la statistique publique

Motivation : vers une base exhaustive sur le patrimoine immobilier des ménages

Une source nouvelle pour des travaux d'études ou de production (voir Courrier des statistiques, janvier 2022) : ▸ Sources

- Constitution d'une base de données exhaustive reliant les patrimoines immobiliers à partir d'une nouvelle exploitation de sources administratives :
  - données fiscales et cadastrales ;
  - données sur les transactions immobilières ;
  - données du registre du commerce et des sociétés ;
- Base de données adossée au fichier Fidéli (composition et localisation des ménages, revenus fiscaux et sociaux).
- Vif intérêt des chercheurs et des acteurs institutionnels (France stratégie, Cnis, SSP, etc.).
- De multiples études et productions rendues possibles.

Plan

Des enjeux méthodologiques importants

Méthode de valorisation

Objectifs des travaux :

- Comparer les modèles hédoniques et de *machine learning*
- Reconstituer finement la distribution du patrimoine immobilier

Difficultés de l'estimation des prix immobiliers :

- Hétérogénéité importante sur l'ensemble du territoire (segmentation des marchés).
- Données géographiques difficiles à prendre en compte dans les modèles économétriques.
- Enjeux aux extrémités :
  - Biens chers souvent sous-évalués ;
  - Biens peu chers souvent sur-évalués ;
  - Zones peu intenses en transactions *vs* zones dynamiques ;
  - Biens fréquemment vendus *vs* biens peu souvent sur le marché ;
- Restriction au champ des logements (à ce stade).

Objectifs de la valorisation et critères de choix

Méthode de valorisation

L'objectif est de définir une méthode de valorisation pour évaluer *au prix de marché* l'ensemble des logements en France métropolitaine, afin d'estimer le patrimoine immobilier des ménages.

Les modèles estimés cherchent à :

- Limiter le biais par rapport aux variables explicatives.
- Être le plus précis possible dans la prédiction.

Critères de choix de modèles :

- Arbitrage biais – variance pouvant être calibré ;
- Être centré géographiquement (biais local) ;
- Être centré en fonction des autres variables d'intérêt (caractéristiques des logements **et** des propriétaires) ;
- Refléter l'hétérogénéité du marché local (variance locale).

Données

Méthode de valorisation

Sources :

1. Données cadastrales appariées au fichier Fidéli : description des logements et caractéristiques des propriétaires ;
2. Données DVF : transactions immobilières sur la période 2015-2019 (hors Alsace-Moselle et Mayotte).

Variables explicatives du prix ▸ Liste complète :

- Sur le bien lui-même : surface, nombre et nature des pièces, date de construction, surface du terrain (maisons uniquement), équipements (eau/gaz/électricité), présence de dépendances...
- Sur la transaction immobilière : date, montant, autres biens vendus avec le logement (garage, parking).
- Sur les propriétaires : niveau de vie, biens possédés, etc.
- Sur la géographie : localisation exacte, zonages Insee (iris, commune, AAV), type de commune (littorale, station touristique), distance à la ville la plus proche, etc.

Méthode de valorisation

Modélisation de la variable expliquée :

- Plusieurs choix possibles pour modéliser le prix d'un logement.
- Résultats présentés ici : logarithme du prix au mètre carré, en écart à la moyenne locale.
- Modélisations séparées pour les maisons et les appartements.

$$\log\left(\frac{P_{ij}}{S_{ij}}\right) - \overline{\log\left(\frac{P}{S}\right)}_j = f(\mathbb{X}_{ij})$$

Calcul de la moyenne locale :

- Quatre niveaux géographiques emboîtés formant une partition du territoire : *IRIS*, commune, EPCI et département.
- Calcul de la moyenne des prix locaux sur la zone la plus fine comprenant un nombre suffisant de transactions ($n_z = 6$ pour l'instant).

Différents types de modèles

Méthode de valorisation

Modèles testés :

- Prédiction = moyenne locale du prix au mètre carré ;
- Régression linéaire (sans interactions) ;
- XGBoost (2000 arbres pour les maisons, 3000 pour les appartements) ;
- Rejetés : *deep learning*, régressions pénalisées.

Données

- Ensemble d'entraînement : 2,6 millions de transactions ;
- Ensemble de test : 600 000 transactions ;
- Données d'extrapolation : 28,3 millions de logements.

Qualité des modèles

Méthode de valorisation

Sur le test	$(log(prix/m^2))$	**Moyenne locale**	**Régression linéaire**	**XGBoost**
R^2	**Maisons**	0.58	0.64	0.69
	Appartements	0.80	0.82	0.85
RMSE	**Maisons**	0.39	0.36	0.33
	Appartements	0.29	0.28	0.25

- Moyenne locale < Régression linéaire < XGBoost

Sur le test	$(prix)$	**Moyenne locale**	**Régression linéaire**	**XGBoost**
R^2	**Maisons**	0.53	0.52	0.58
	Appartements	0.85	0.85	0.87
RMSE	**Maisons**	174 861	221 976	165 450
	Appartements	87 750	85 506	79 167

- Maisons < Appartements
- Un ordre de prédiction satisfaisant ▸ Prédictions

Erreur relative de prédiction

Méthode de valorisation

$$\text{Erreur relative par logement : } \hat{\delta}_i = \frac{\hat{P}_i - P_i}{P_i}$$

Table – Performance des modèles : erreurs relatives

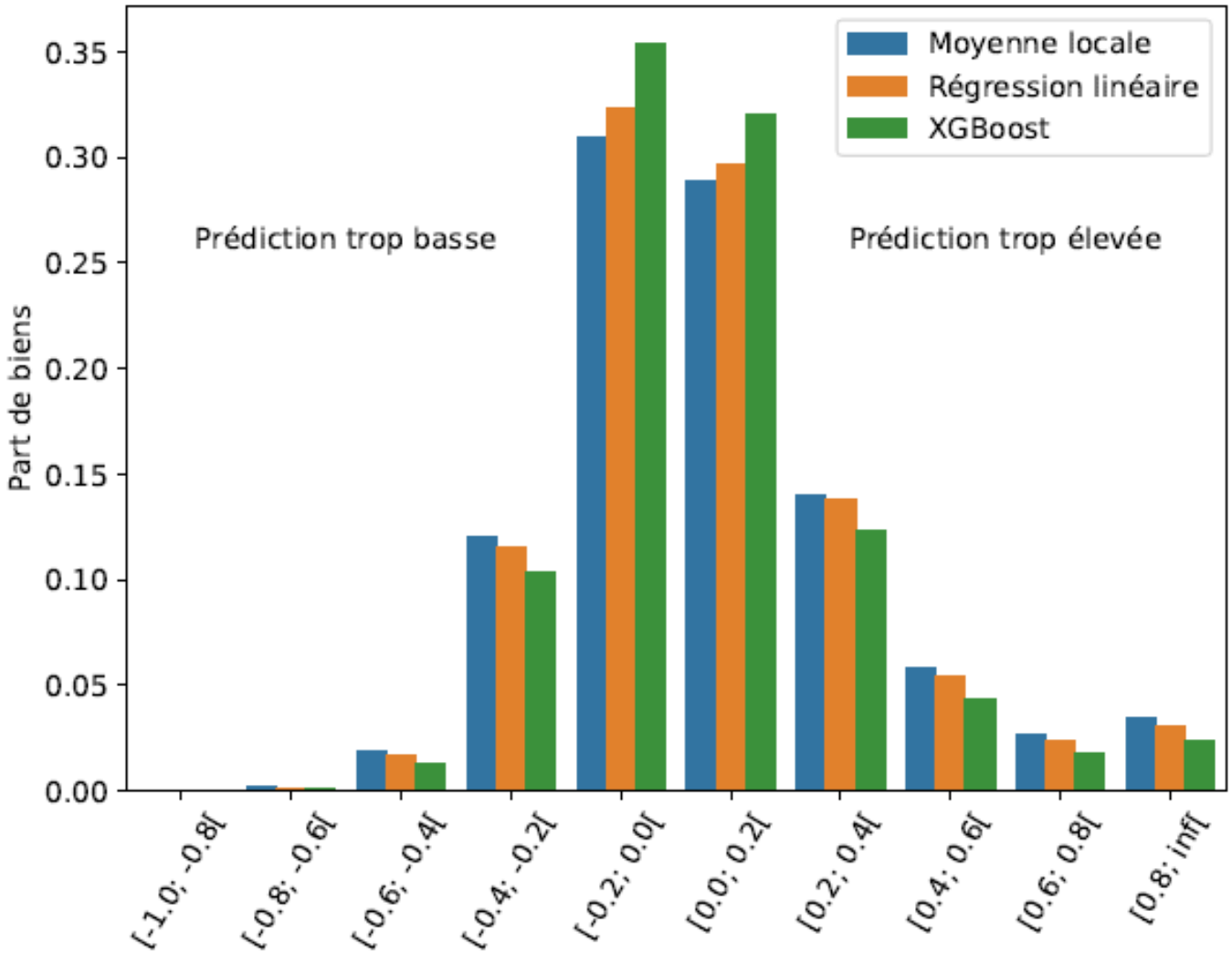
Modèle	Moyenne	Moyenne par quintile de prix				
		Q1	Q2	Q3	Q4	Q5
Maisons						
Moyenne locale	18,2 %	73,9 %	13,6 %	3,1 %	1,1 %	-0,4 %
Régression Linéaire	14,5 %	64,6 %	9,2 %	1,2 %	-0,2 %	-2,1 %
Boosting	11,2 %	51,6 %	8,6 %	1,8 %	-0,9 %	-4,9 %
Appartements						
Moyenne locale	9,1 %	27,5 %	11,7 %	6,1 %	3,1 %	-3,2 %
Régression Linéaire	7,7 %	27,8 %	10,2 %	4,4 %	1,6 %	-5,5 %
Boosting	5,5 %	22,0 %	7,8 %	2,4 %	-0,6 %	-4,1 %

Note : Tous les indicateurs sont calculés sur l'ensemble de test.

- Entre modèles ; Extrémités ; Maisons / appartements

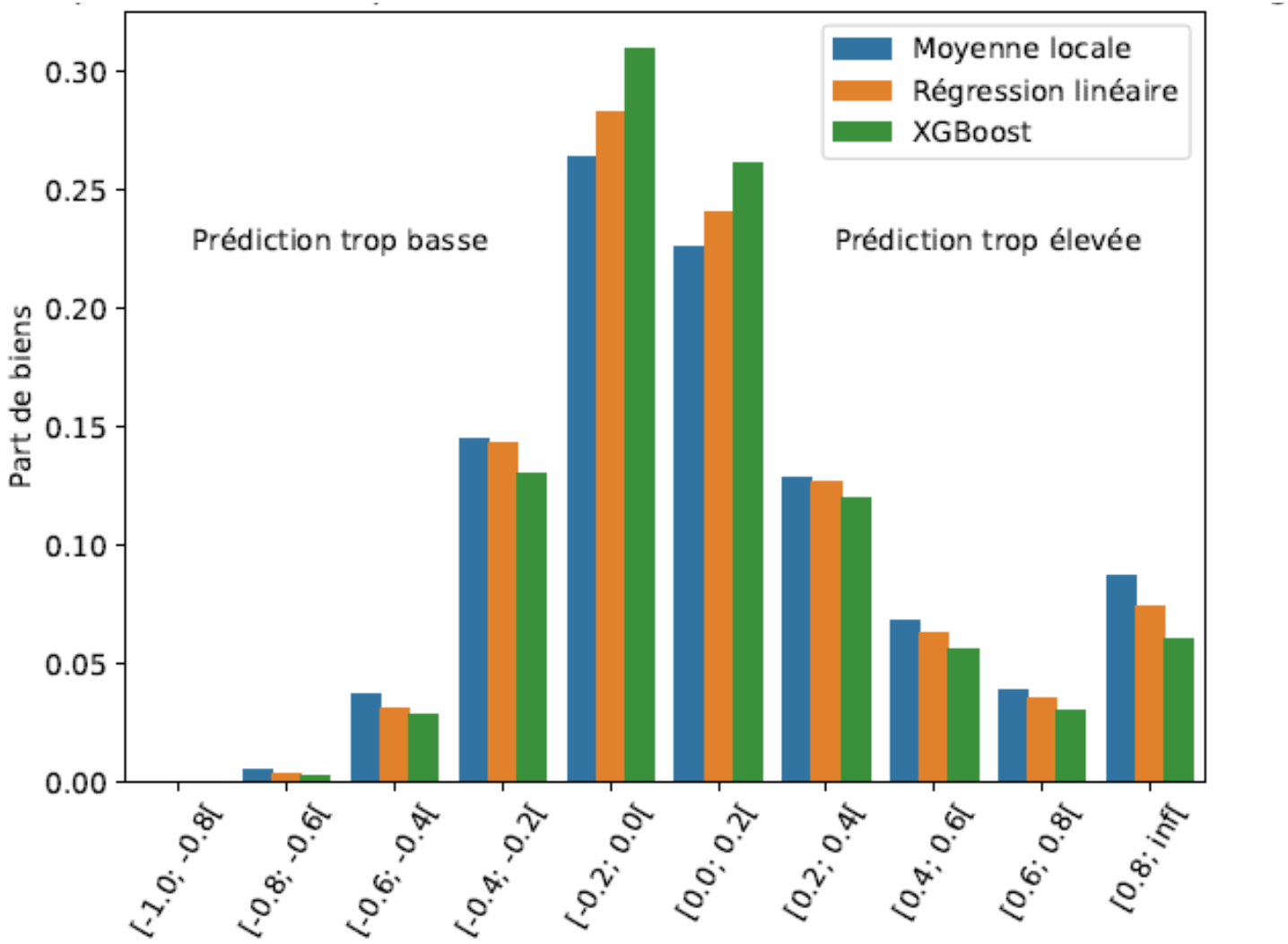
Écart relatif de prédiction (appartements)

Méthode de valorisation



Écart relatif de prédiction (maisons)

Méthode de valorisation



Écart relatif de prédiction

Méthode de valorisation

Table – Part des erreurs supérieures à 20 % en valeur absolue

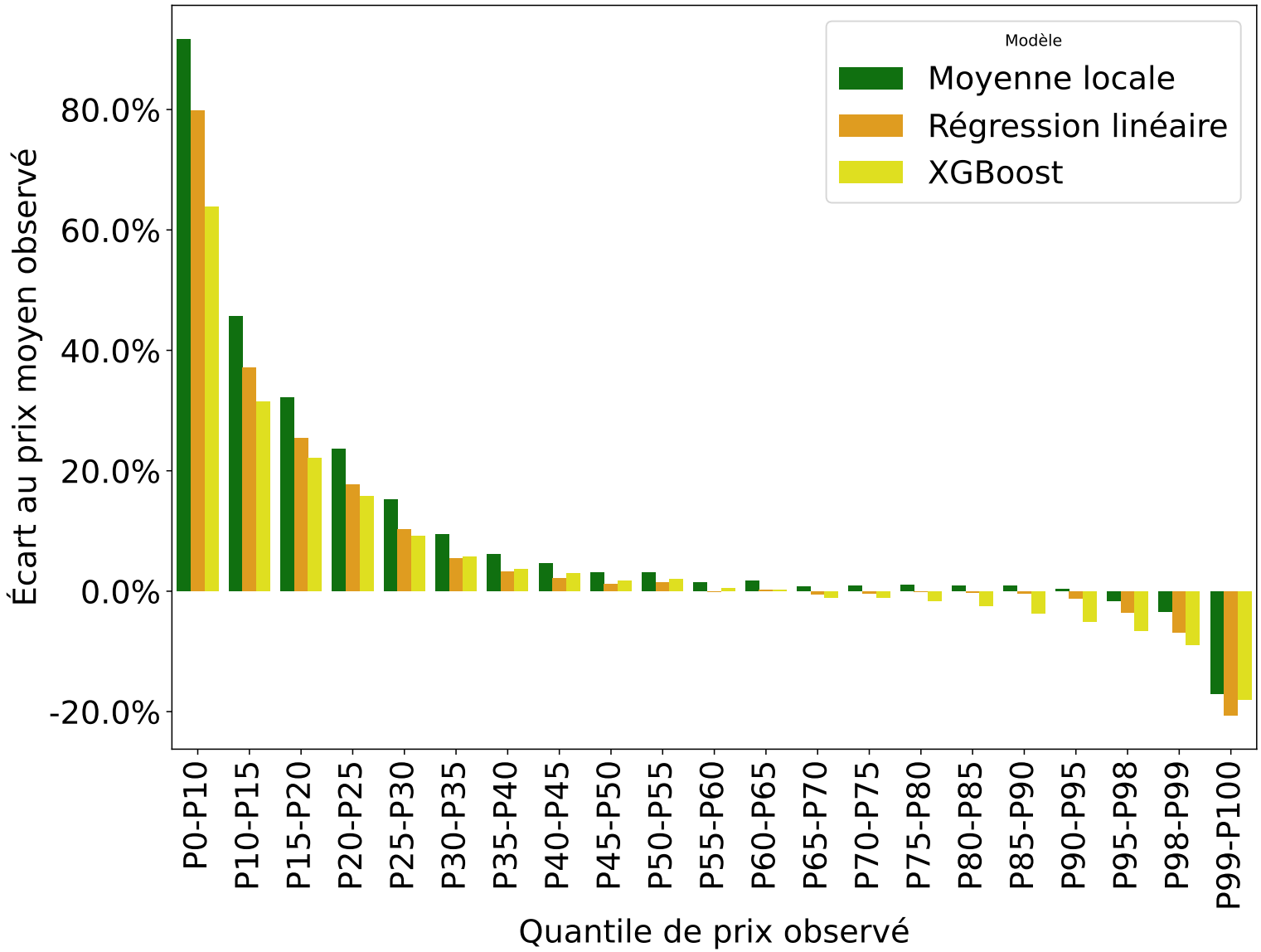
Modèle	Part moyenne	Part par quintile de prix				
		Q1	Q2	Q3	Q4	Q5
Maisons						
Moyenne locale	52,0 %	75,1 %	52,9 %	45,5 %	42,1 %	44,3 %
Régression Linéaire	46,7 %	72,3 %	45,2 %	38,3 %	36,1 %	41,4 %
Boosting	39,4 %	68,3 %	41,7 %	30,3 %	26,3 %	30,5 %
Appartements						
Moyenne locale	40,4 %	53,7 %	43,9 %	38,3 %	34,5 %	31,4 %
Régression Linéaire	36,2 %	51,6 %	38,7 %	32,3 %	29,4 %	28,8 %
Boosting	28,4 %	43,3 %	30,6 %	23,0 %	21,4 %	23,4 %

Note : Tous les indicateurs sont calculés sur l'ensemble de test.

- Entre modèles ; Extrémités

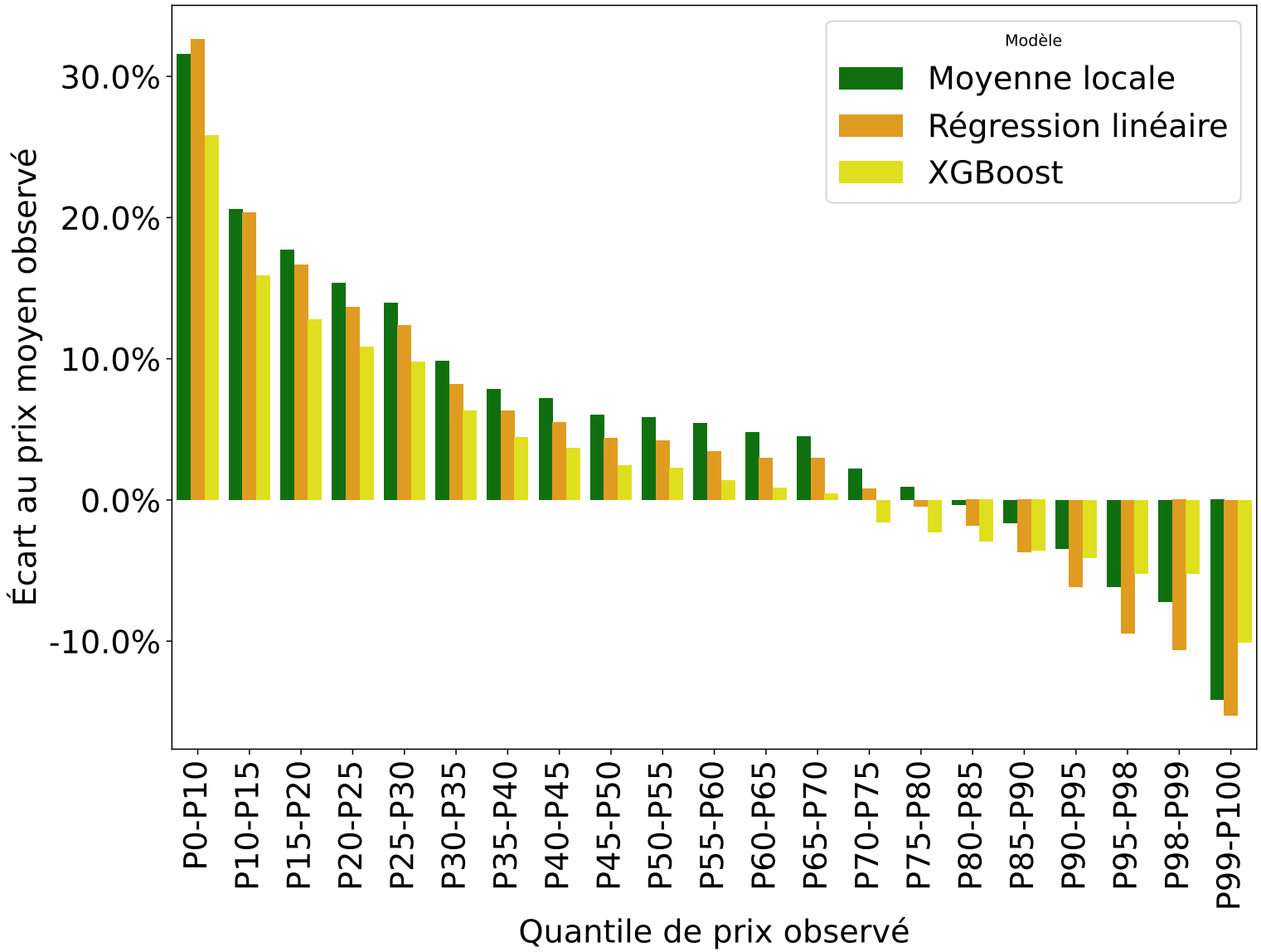
Résultats : les biens extrêmes sont mal valorisés

Maisons



Résultats : les biens extrêmes sont mal valorisés

Appartements



Résultats : reproduction des montants totaux de transactions

Méthode de valorisation

Table – Reproduction des montants totaux de transaction (2015-2019)

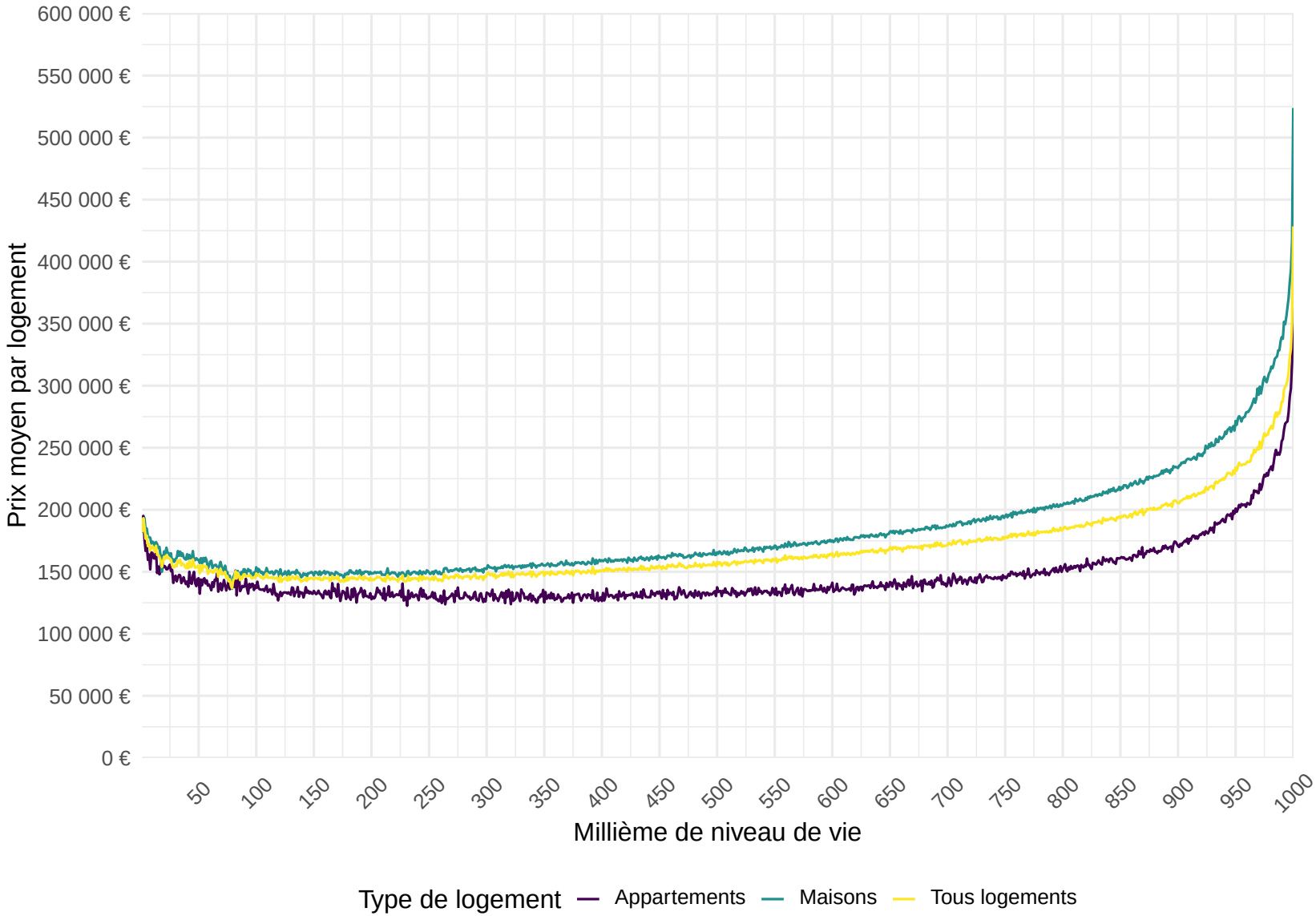
	Montants observés, en Md€	Ratio total prédit/total observé		
		Moyenne locale	Régression linéaire	Boosting
Maisons	**385,24**	**104,37 %**	**101,89 %**	**100,30 %**
Appartements	**263,37**	**101,37 %**	**99,50 %**	**99,48 %**
Tous logements, statifiés par taille d'aire d'attraction des villes				
Communes isolées	32,57	100,89 %	97,47 %	98,15 %
Moins de 50 000 habitants	58,55	100,70 %	98,05 %	98,64 %
50 000-200 000 habitants	81,15	102,25 %	100,06 %	99,52 %
200 000-700 000 habitants	130,88	102,77 %	100,52 %	99,73 %
700 000 habitants ou plus	136,48	104,05 %	102,70 %	100,54 %
Aire de Paris	208,98	104,20 %	101,68 %	100,58 %
Ensemble	**648,61**	**103,15 %**	**100,92 %**	**99,97 %**

Note : Tous les ratios total prédit/total observé sont calculés sur l'ensemble de test.

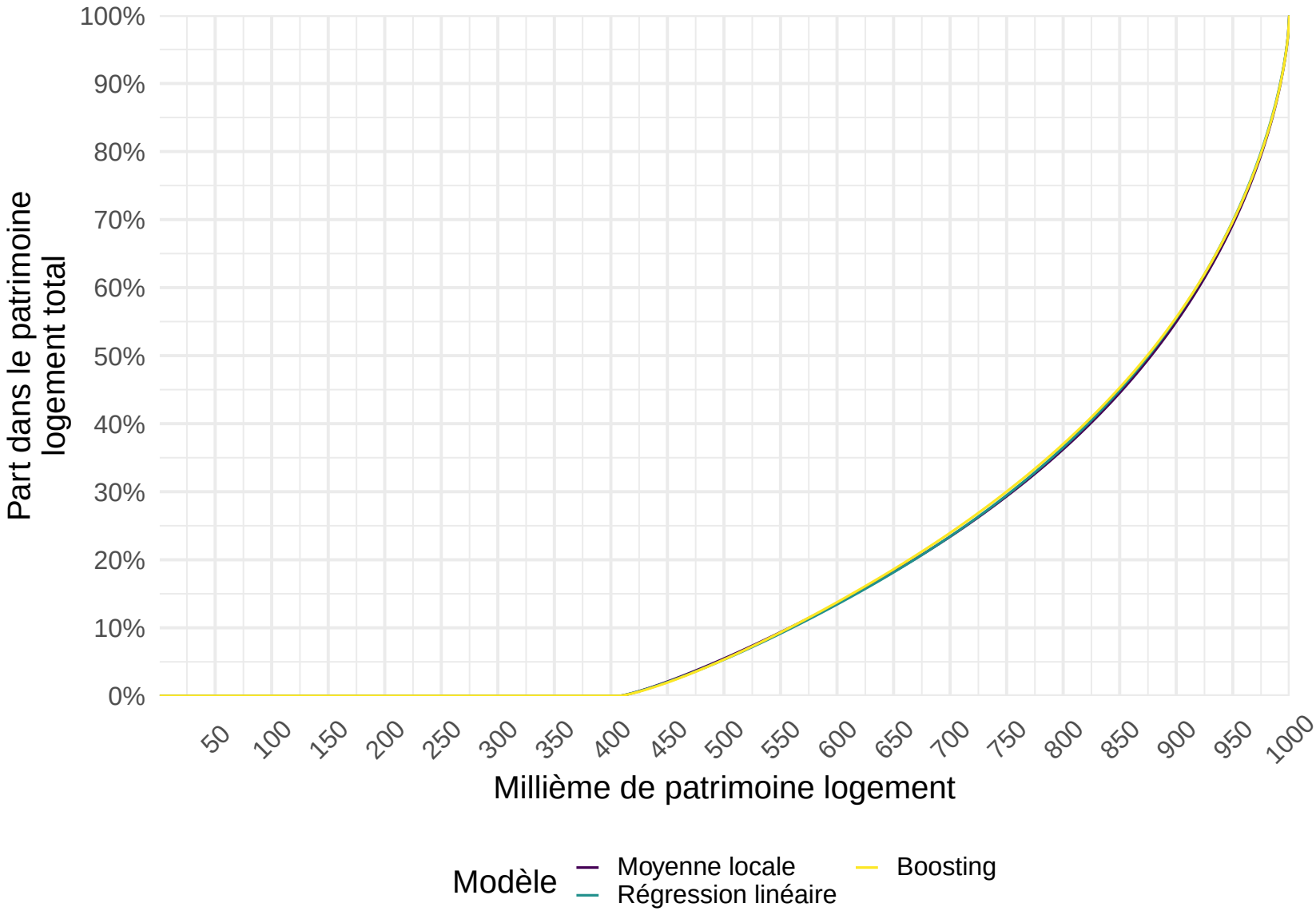
Plan

Prix moyen des logements et niveau de vie

Le prix moyen estimé augmente avec le niveau de vie



Une concentration marquée, identique selon les modèles



Une distribution peu sensible au modèle utilisé

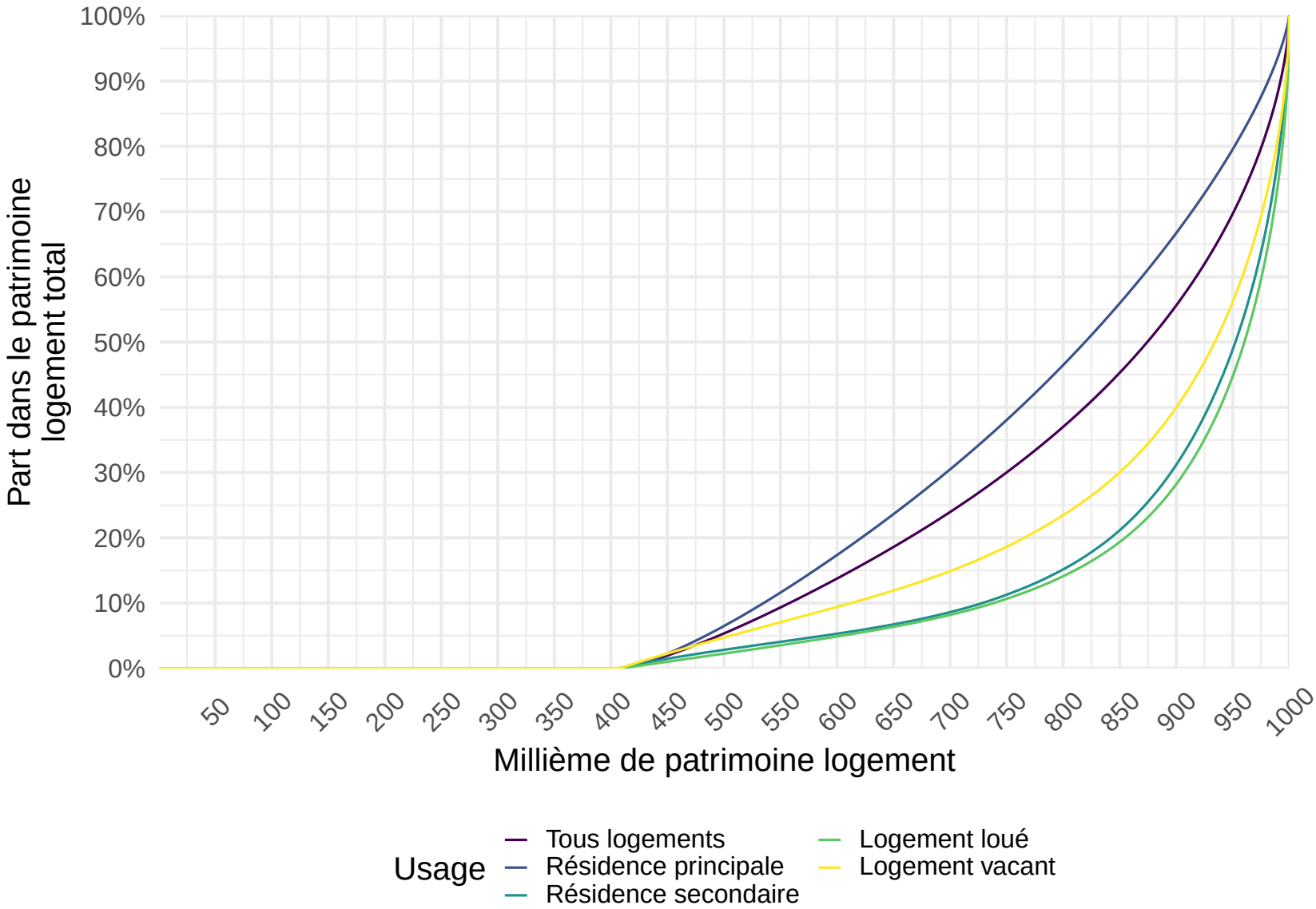
Table – Patrimoine immobilier brut des ménages possédant au moins un logement, évalué au premier trimestre 2017

Statistique	Régression linéaire	Boosting
P10	88,9 k€	82,7 k€
P25	129,4 k€	125,4 k€
Moyenne	300,0 k€	294,5 k€
Médiane	201,7 k€	198,9 k€
P75	340,7 k€	337,5 k€
P90	584,8 k€	578,4 k€
P95	827,8 k€	817,4 k€
P99	1 643,5 k€	1 616,4 k€
P99,5	2 144,5 k€	2 105,0 k€
P99,9	3 937,2 k€	3 823,9 k€

Note : Tous les indicateurs sont calculés sur l'ensemble de test.

Concentration du patrimoine et usage des logements

Le patrimoine locatif est la composante la plus concentrée



Des premiers résultats sur le prix et le patrimoine **brut**

Distributions selon le niveau de vie

Augmentation importante avec le niveau du vie des propriétaires :

- du prix au mètre carré des appartements (et légère pour les maisons) ▸ Prix au m^2
- de la surface des maisons (alors que celle des appartements est stable) ▸ Surface
- de la part de patrimoine possédée sous la forme d'appartements (et lègèrement pour les maisons) ▸ Patrimoine

Le patrimoine immobilier est plus concentré que le revenu mais moins que le patrimoine financier (HVP) : ▸ Concentration en montants

- moyenne : 300 000 € ; médiane : 200 000 €
- P10 : 85 000 € ; P25 : 130 000 €
- P99 : 1,6 million € ; P99,9 : 3,9 millions €

Plan

Conclusion

Résultats préliminaires

Bilan méthodologique provisoire :

- L'algorithme de *boosting* présente de meilleures performances que les autres approches...
- ... mais la moyenne locale est déjà un bon estimateur moyen !
- Conditionnellement aux caractéristiques des logements, les prix prédits sont proches des prix observés en moyenne ;
- La prédiction du prix au niveau de chaque bien présente une variance résiduelle élevée et est à améliorer ;
- Les différentes approches peinent à capter les queues de la distribution.

Bilan économique provisoire :

- La distribution du patrimoine immobilier est peu sensible au modèle retenu.

Travail en cours : perspectives

Amélioration de la structure de prédiction

- Plusieurs modèles séparés par zones
- *Matching* sur les biens extrêmes

Informations explicatives complémentaires

- Données géographiques plus fines : temps de transport et proximité des services publics, qualité de l'environnement (pollution, paysages, météo, etc.).
- Caractéristiques socio-économiques de la zone : structure par âge et par niveau de la population locale.
- Informations sur les logements : taxe foncière, DPE...

Comparaison aux sources existantes

- Comptabilité nationale.
- Enquête Histoire de Vie et Patrimoine.

Plan

Cinq sources principales de données

Une base exhaustive du patrimoine immobilier

- Les données cadastrales (fichiers Majic) ;
  - Description des propriétés bâties (maison, appartement, garage...).
  - Description des propriétés non bâties (jardins, champs, vergers...).
  - Identité des propriétaires (état civil, adresse, nature du droit de propriété).
  - Caractéristiques fiscales du logement (valeur locative et exonérations).
- Le fichier Fidéli ;
  - Description des logements (surface, localisation, ascenseur...).
  - Description des ménages vivant dans ces logements (état civil des personnes, revenus).
  - Identification des individus résidents en France.
- Les données du registre du commerce et des sociétés (RCS) ;
  - Informations sur les sociétés (forme juridique, adresse du siège) et sur les personnes physiques qui en sont représentantes (gérants, actionnaires, associés...).
- Les données sur les transactions immobilières (DVF) ;
  - Descriptions des transactions immobilières sur la période 2015-2019 : nature et caractéristiques des biens vendus, montant et date de la transaction. Données non disponibles sur l'Alsace-Moselle et Mayotte.
- Les données sur la fiscalité locale (REI).
  - Taux des impôts locaux par collectivité locale.

Retour

Rappels sur XGBoost

Arbres de décision et forêts aléatoires

1. On tire un sous-échantillon de colonnes et de lignes.
2. On coupe en deux pour optimiser le *CART-criterion* (plus grand écart entre la variance actuelle et la variance post-découpage).
3. On recommence jusqu'à construire un arbre : À chaque cellule on associe la valeur moyenne de la *target*.
4. On fait un grand nombre d'arbres.

XGBoost : *eXtreme Gradient Boosting*

- Succession de plusieurs modèles qui apprennent à prédire :
  1. y : Résidus ε_1
  2. Puis ε_1 : Résidus ε_2
  3. Etc.
- Intérêt : Donne plus de poids aux lignes difficiles à prédire (car leur résidu est plus grand).

Nettoyage des données

Etape 1 : *Clips* sur les très gros biens :

Caractéristique	Valeur minimale	Valeur maximale
Etage	-1	10
Nb pièces	0	22
Chambres	0	9
Salles à manger	0	5
Cuisines $< 8m^2$	0	2
Cuisines autre	0	2
Nb pièces annexes	0	9
Baignoires	0	3
Douches	0	4
Lavabos	0	8
WC	0	5
Salles d'eau	0	6
Garages	0	3
Piscines	0	1
Caves & greniers	0	5
Terrasses	0	2
Autres dépendances	0	2
Surface	$15m^2$	$600m^2$
Montant	24 000	$\emptyset$

Retour

Nettoyage des données

Etape 2 : On supprime les endroits où on manque de données :

- Alsace-Moselle
- DOM/TOM : '973', '976', '971','972','974','977' & '97127', '29155', '85113', '22016'

Etape 3 : Les erreurs de prix ou de surface

On considère qu'un bien est une erreur s'il vérifie les deux conditions :

- Son prix au mètre carré est dans le top 0.1%.
- Il y a moins de 3 biens dans sa commune qui ont un prix au mètre carré dans le top 0.5%.

Ou s'il vérifie l'une des deux conditions :

- Son prix au mètre carré est inférieur à 10 fois le prix au mètre carré médian local.
- Son prix au mètre carré est dans le bottom 1%.

Liste variables

Caractéristiques du bien

- Type de bien
- Eau, électricité, gaz, ascenseur, escalier de service, chauffage central, vide ordure, tout à l'égout
- Etage
- Période de construction
- Quartier prioritaire de la ville
- HLM
- Pièces, chambres, salle de bain, ...
- Surface agricole, au sol, bois, lac
- Cave, grenier, piscine, garage, autre dépendance
- log(Surface), au carré, au cube

Liste variables

Géographie

- Indicateurs locaux
- Type de zone locale (*IRIS*, commune,...)
- Littoral
- Type de station touristique
- Distance à la plus proche de ville de 50 000, 100 000, 200 000 et 500 000 habitants

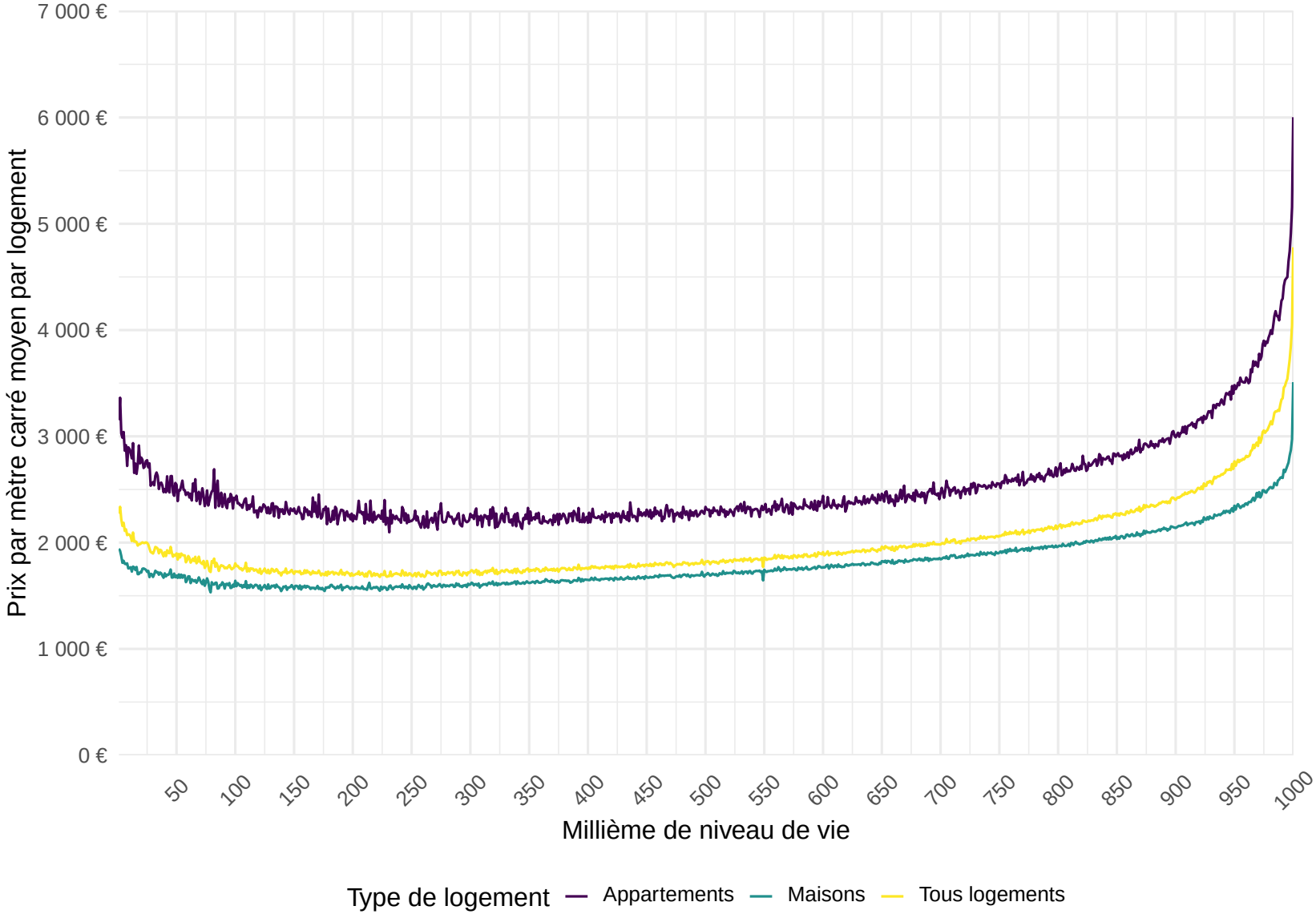
Sur la mutation

- Année et trimestre de vente

Retour

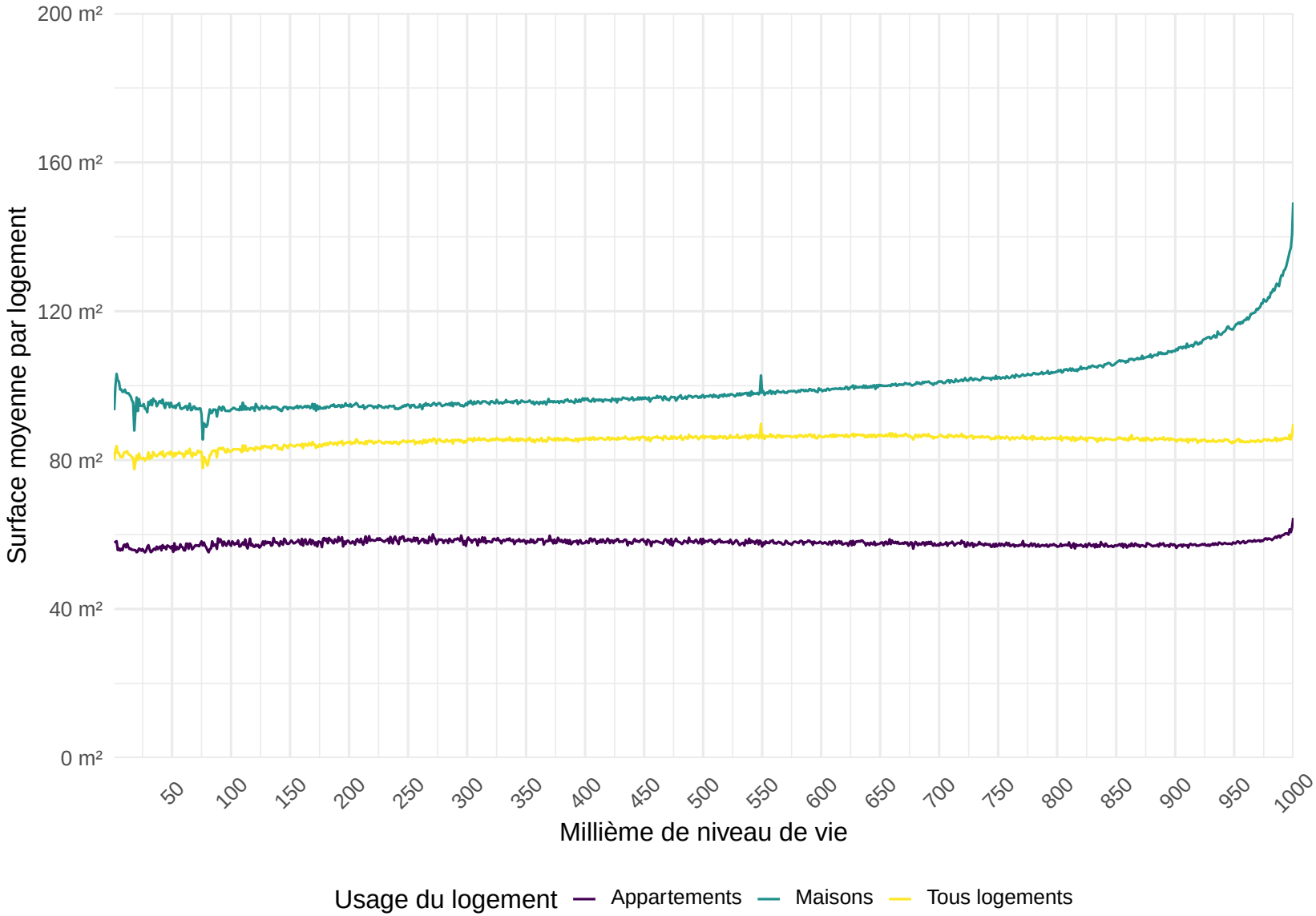
Prix au mètre carré et niveau de vie

Augmentation importante du prix au mètre carré des appartements



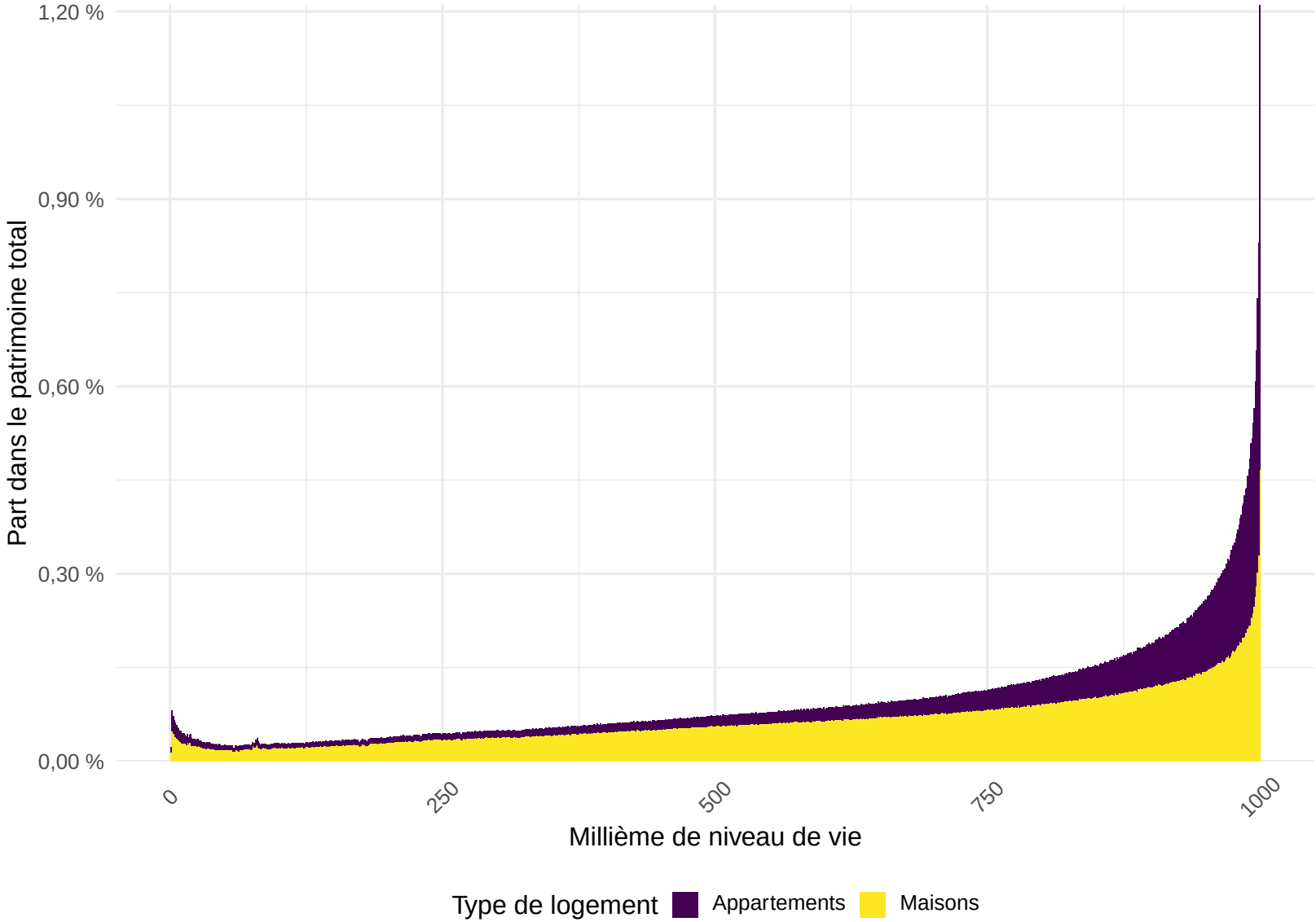
Surface et niveau de vie

Augmentation importante de la surface des maisons



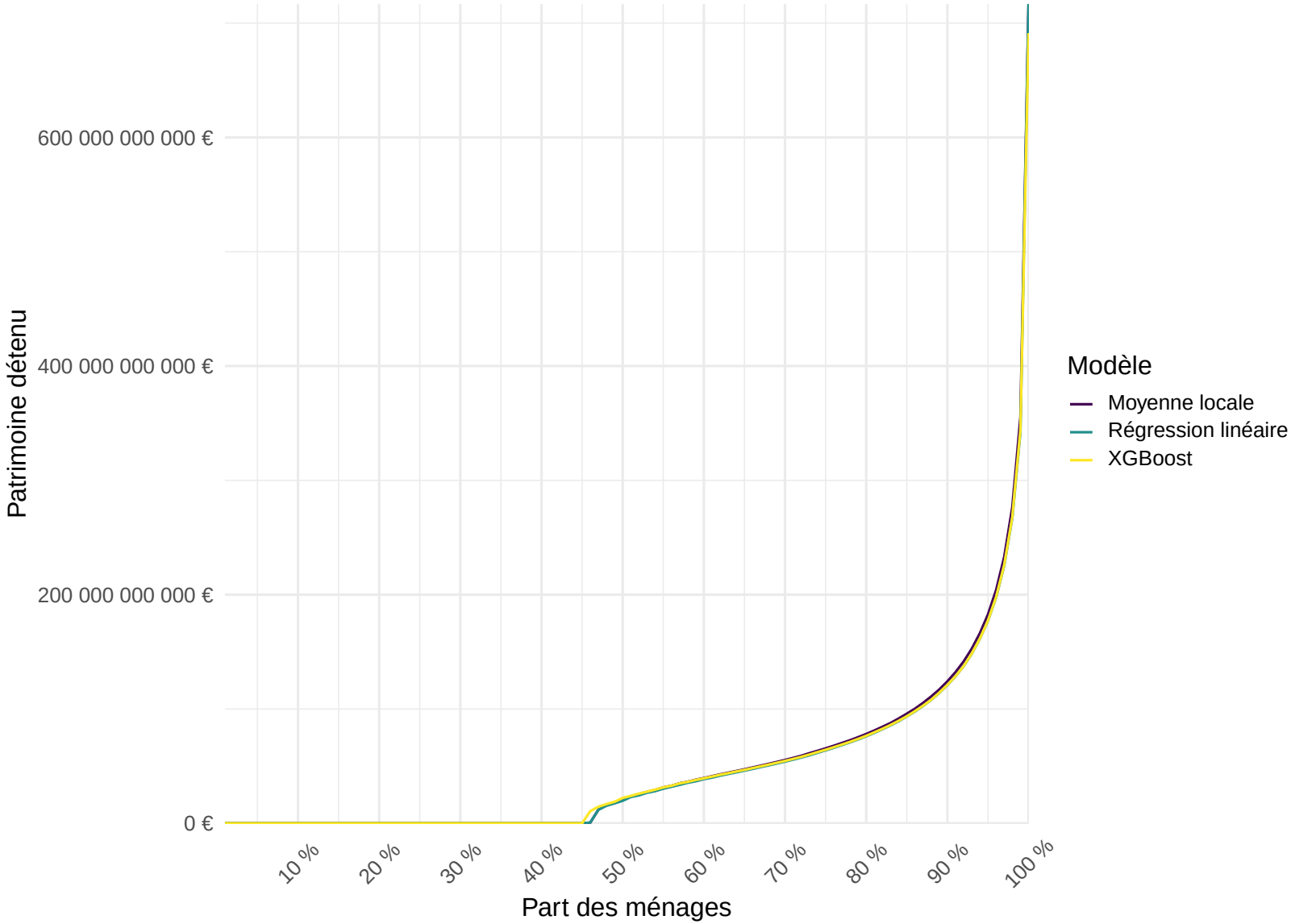
Répartition du patrimoine et niveau de vie

La répartition des appartements est plus concentrée



Une concentration marquée, identique selon les modèles

Répartition du patrimoine immobilier (en montants)



L'ordre de prédiction

▸ Retour Modèles ▸ Retour Distributions

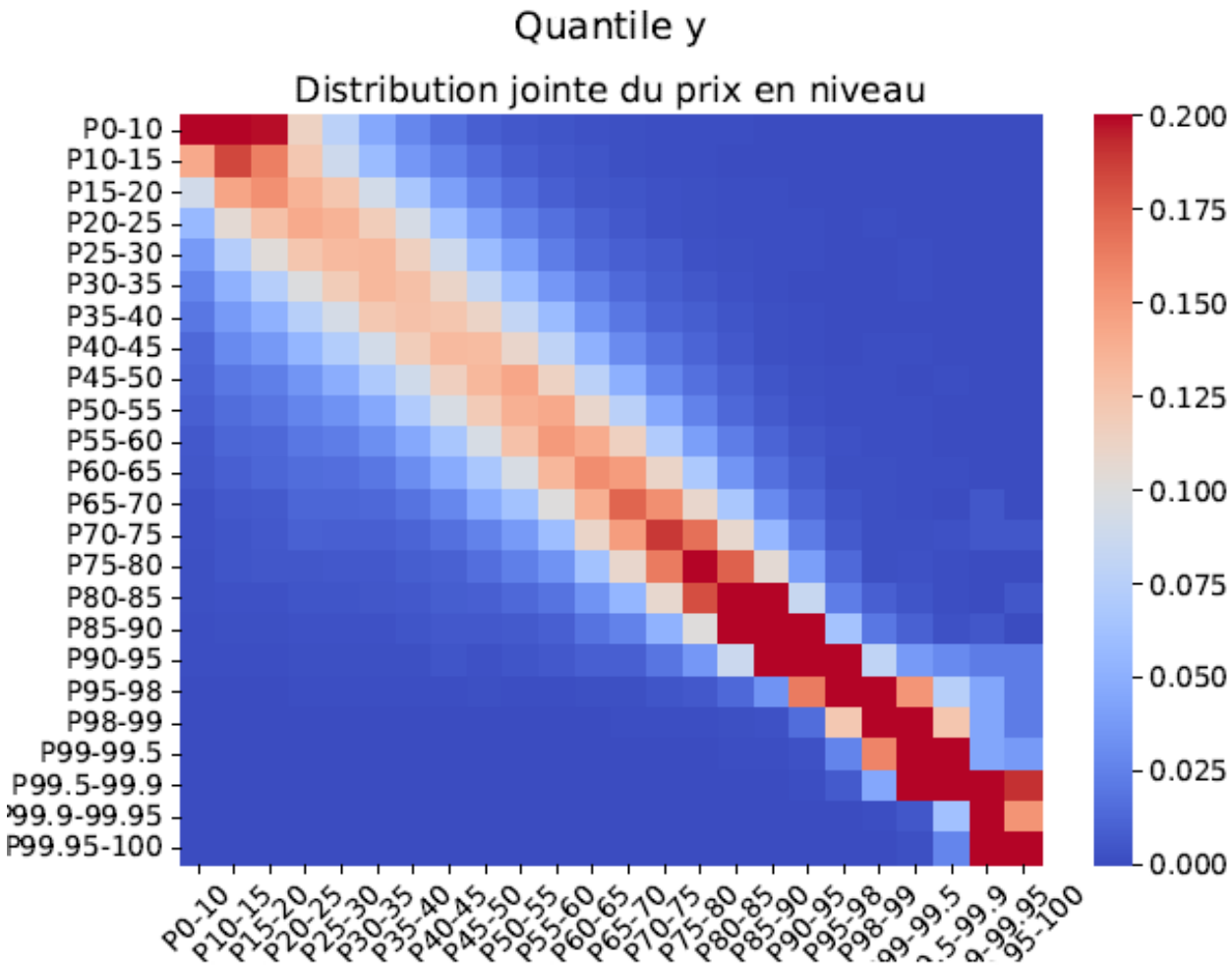


Figure – Distributions jointes de probabilité sur les maisons.