Modélisation du prix de l'Immobilier à partir de données géographiques et structurelles

Le Roux Noa - 20 août 2025
Master 2 Econométrie et Statistique, parcours Econométrie Appliquée

Cette étude analyse l'impact du Diagnostic de Performance Énergétique (DPE) sur les prix immobiliers à Brest, en mobilisant à la fois l'économétrie spatiale et le machine learning. Les résultats montrent que le marché valorise les logements performants (classes B et C) et sanctionne fortement les passoires thermiques (F et G), avec des effets diffusés à l'échelle des quartiers. L'économétrie spatiale met en évidence ces tendances globales et l'importance des interactions de voisinage. Le machine learning, quant à lui, confirme le rôle du DPE mais souligne que d'autres variables comme la surface ou la localisation dominent dans la hiérarchie globale des prix. Son apport majeur réside dans l'explicabilité locale grâce aux Shapley Values, permettant d'identifier, logement par logement, les facteurs déterminants du prix. Ces résultats ouvrent la voie à un outil de scoring énergétique utile pour Arkéa, combinant gestion du risque et accompagnement des clients dans la transition énergétique.

Marché immobilier breton, DPE, Prix de l'immobilier,Brest, Économétrie spatiale, Modèle SEM, SAR, SLX, SDM, Machine-Learning, Explicabilité

1 Cadre de l'étude

Le marché immobilier breton, à l'image du reste du territoire national, connaît depuis l'été 2022 une phase d'ajustement caractérisée par une contraction des volumes de transactions et une baisse des prix. Dans ce contexte de rééquilibrage, l'influence du DPE est devenue un critère structurant, redéfinissant les attentes des acquéreurs et la valeur des biens. Un logement classé F ou G, qualifié de "passoire thermique", subit une décote moyenne de 15% sur son prix de vente et un délai de vente plus long, une dynamique directement liée à la loi Climat et Résilience. Cette loi introduit des interdictions progressives de location pour les logements les moins performants, avec des échéances déjà en vigueur depuis janvier 2023 pour les logements classés G+ et janvier 2025 pour tous les logements classés G. Ces régulations incitent de nombreux propriétaires à retirer leurs biens du marché locatif, modifiant l'offre disponible.

Pour le Crédit Mutuel Arkéa, comprendre ces dynamiques est essentiel afin d'anticiper les évolutions du marché, d'adapter ses politiques de financement immobilier et d'évaluer les risques associés aux biens dévalorisés ou non conformes aux exigences réglementaires. Cette étude s'inscrit précisément dans une demande de la direction des risques du groupe Arkéa visant à évaluer l'impact de l'étiquette DPE sur les prix de vente immobiliers, notamment depuis l'entrée en vigueur de la réglementation DPE 2025. Une analyse initiale à l'échelle de la Bretagne, utilisant un modèle de régression linéaire multiple, avait déjà examiné ces impacts. Cependant, l'ampleur du jeu de données (environ 68 000 observa-

tions) a rendu l'application de modèles d'économétrie spatiale, plus exigeants en calcul, techniquement inenvisageable à cette échelle.

Afin d'affiner l'analyse et de permettre une modélisation spatiale rigoureuse, l'étude a été restreinte à un sous-échantillon plus maniable, en se concentrant exclusivement sur le périmètre de la ville de **Brest**. Dans ce cadre, la problématique centrale de cette note est de présenter et d'interpréter les résultats d'une analyse en économétrie spatiale sur les déterminants du prix de vente des maisons et appartements à Brest. L'objectif est de mettre en lumière l'influence du DPE et d'autres caractéristiques, tout en tenant compte des interactions spatiales entre les biens immobiliers. Cette approche vise à offrir une compréhension plus fine des facteurs de prix, tant de manière globale (*modèles d'économétrie spatiale*) qu'individuelle (*algorithme d'explicabilité de machine-learning*), et à poser les bases de réflexions stratégiques pour le Crédit Mutuel Arkéa.

2 Méthodologie : L'Économétrie Spatiale pour une Analyse Approfondie

Pour saisir les dynamiques complexes du marché immobilier de Brest, une méthodologie d'économétrie spatiale a été privilégiée, permettant de modéliser les prix des biens (appartements et maisons) en tenant compte des interactions potentielles entre propriétés voisines. Cette approche est cruciale car elle intègre deux dimensions fondamentales : **l'hétérogénéité spatiale**, qui reflète la variabilité des comportements selon la localisation géographique, et **l'autocorrélation spatiale**, qui mesure le degré d'interdépendance entre unités spatiales.

Le choix de la matrice de pondération spatiale, qui définit le voisinage de chaque bâtiment, est une étape déterminante. Contrairement aux unités spatiales classiques comme les départements ou les régions, les polygones représentant des bâtiments ne sont généralement pas contigus, étant souvent séparés par des espaces publics. Par conséquent, les matrices de contiguïté de type Reine ou Tour se sont avérées inappropriées. Il a été plus judicieux de recourir à des matrices basées sur les k **P**lus **P**roches **V**oisins (PPV), définissant les relations spatiales à partir de la distance euclidienne entre les centroïdes des polygones. Les matrices à 1 (PPV1) et 3 (PPV3) plus proches voisins ont été utilisées, offrant une meilleure capture des interactions de proximité dans un environnement urbain dense.

Avant la modélisation, l'existence d'une autocorrélation spatiale significative des prix a été confirmée par l'indice de Moran. Pour les appartements, les indices de Moran étaient de 0,35 (PPV1) et 0,30 (PPV3), tous deux hautement significatifs, révélant un fort regroupement géographique des prix similaires. Pour les maisons, l'autocorrélation était également positive et significative, bien que légèrement moins marquée (I = 0,26 pour PPV1 et 0,21 pour PPV3). Ces résultats ont permis de rejeter l'hypothèse d'indépendance spatiale, justifiant pleinement l'adoption d'un modèle spatial.

La sélection du modèle économétrique spatial le plus approprié (**SAR** : *Simultaneous AutoRegressive Model* , **SEM** : *Spatial Error Model*, **SDM** : *Spatial Durbin Model* ou **SLX** : *Spatially Lagged X*) a été effectuée à l'aide de tests de Moran sur les résidus d'une régression linéaire ordinaire (OLS) et de tests de Lagrange Multiplier (LMerr, LMLag) ainsi que leurs versions robustes (RLMerr, RLMLag). Pour les appartements, les tests ont d'abord suggéré un modèle SAR pour la matrice PPV1 et un modèle SEM pour la matrice PPV3. Cependant, des tests de rapport de vraisemblance (LR) et les critères d'information d'Akaike (AIC) ont ensuite indiqué que, pour la matrice PPV3, le Modèle SDM offrait un ajustement statistiquement meilleur que le SEM, justifiant ainsi son adoption. Pour la matrice PPV1, le SAR restait privilégié face au SDM. Pour les maisons, la rigueur de la sélection a également conduit à privilégier le SDM pour les deux matrices PPV1 et PPV3, au vu de sa supériorité significative révélée par les tests LR et les valeurs d'AIC. Le SDM est un modèle complet qui inclut à la

fois la variable dépendante et les variables explicatives spatialement décalées, permettant de capter les effets directs et indirects.

3 Résultats et Interprétation des Modèles Spatiaux

L'interprétation des modèles spatiaux, notamment le SDM, est complexe car les coefficients ne peuvent être directement lus comme des effets marginaux unitaires en raison de l'autocorrélation spatiale. Il est impératif d'analyser les effets **directs** (impact sur le bien lui-même), **indirects** (effets de débordement sur les biens voisins) et **totaux** (somme des deux).

L'analyse des modèles SDM retenus pour les appartements (matrice PPV3) et les maisons (matrice PPV1) a révélé une autocorrélation spatiale positive et significative des prix dans les deux cas ($\rho = 0.1454$ pour les appartements, $\rho = 0.0651$ pour les maisons). Cela confirme que les valeurs immobilières tendent à se regrouper géographiquement, les propriétés ayant des prix similaires se situant souvent à proximité les unes des autres.

Concernant le DPE, les résultats sont convergents et significatifs. Relativement à la classe D (référence), les biens classés B et C sont systématiquement associés à une valorisation positive des prix pour les appartements et les maisons, reflétant une "prime énergétique". En revanche, les classes E, F et G entraînent une décote significative, confirmant que les "passoires thermiques" sont pénalisées sur le marché. Pour les appartements, la classe A a montré un effet total négatif (-1,5882), probablement dû à sa rareté dans l'échantillon, tandis que pour les maisons, elle affichait un effet positif, bien que modéré.

Les caractéristiques physiques des biens confirment leur rôle prépondérant dans la détermination des prix. L'âge du bâtiment exerce un effet négatif significatif, les constructions plus anciennes étant moins valorisées. À l'inverse, la surface habitable et le nombre de pièces contribuent positivement et significativement aux prix pour les deux types de biens, confirmant la valorisation des espaces plus grands et mieux agencés. La surface du terrain est également un facteur valorisant pour les maisons, bien que son impact marginal diminue au-delà de certaines tailles.

L'influence des variables géographiques est complexe. La distance aux entités "bleues" (océan, merc, lac) est systématiquement associée à une valorisation de la proximité, un éloignement entraînant une diminution des prix. Inversement, l'éloignement des entités "rouges" (nuisances) est généralement apprécié, suggérant une prime pour la tranquillité. Les effets des entités "vertes" (espaces végétaux) sont plus nuancés, montrant parfois des effets directs négatifs (la proximité étant valorisée) compensés par des effets indirects positifs, ce qui peut suggérer des dynamiques spatiales complexes de substitution ou de composition. Ces résultats soulignent que la localisation n'est pas qu'une question de proximité directe, mais aussi d'influence du voisinage.

4 Complémentarité avec le Machine Learning

L'intégration du Machine Learning constitue une étape centrale de l'analyse, destinée à affiner l'étude des déterminants du prix de vente des biens immobiliers après l'application des modèles d'économétrie spatiale. Cette approche mobilise des techniques de modélisation avancées, en particulier l'algorithme Random Forest, associé à des outils d'explicabilité, afin de mieux saisir les facteurs qui influencent les prix, aussi bien à une échelle générale (tendances globales) qu'individuelle (cas spécifiques).

Le Random Forest se distingue par sa capacité à capturer des relations complexes, non linéaires et interactives entre variables, tout en s'affranchissant des hypothèses distributionnelles strictes des modèles linéaires classiques. Sa robustesse repose sur l'agrégation des prédictions issues de multiples arbres de décision, tandis

que l'optimisation de ses hyperparamètres par recherche aléatoire et validation croisée permet d'accroître sa performance prédictive.

Pour pallier l'opacité de ces modèles, différentes méthodes d'interprétabilité sont mobilisées. Les Partial Dependence Plots (PDP) mettent en évidence les effets marginaux moyens des variables, tels qu'une relation quasi linéaire croissante entre la surface habitable et le prix, ou encore une décroissance de la valeur à mesure que l'on s'éloigne des zones littorales ou lacustres. L'importance par permutation permet, quant à elle, d'évaluer la contribution effective de chaque variable à la qualité prédictive du modèle. Toutefois, l'apport le plus significatif réside dans l'analyse locale des prédictions, rendue possible par les Shapley Values, qui attribuent à chaque caractéristique sa part de responsabilité dans la valeur prédite pour un bien donné.

Variable	Shapley Values
DIST_VERT_KM	-0.16
DIST_ROUGE_KM	-0.07
surface_habitable	-0.03
surface_terrain	-0.03
nb_pieces	-0.01
code_dep	0.00
DPE_score	+0.03
nb_etages	+0.03
age_batiment	+0.05
DIST_BLEU_KM	+0.06

TABLE 1 – Shapley Values pour le bien n°50

Ainsi, pour le bien n° 50, le modèle estime un logarithme du prix de 12.001, légèrement inférieur à la moyenne attendue (12.166). Les Shapley Values révèlent que l'éloignement d'environ un kilomètre des zones littorales ou lacustres (variable DIST BLEU KM) exerce l'effet négatif le plus marqué (-0.16). Le code départemental contribue également à une baisse notable (-0.07), traduisant un effet territorial défavorable. À l'inverse, certaines caractéristiques jouent positivement sur la prédiction : le nombre de pièces (+0.06), la surface du terrain (+0.05) et, dans une moindre mesure, la surface habitable (+0.03), reflètent un profil de logement spacieux valorisé par le modèle. D'autres variables, comme l'âge du bâtiment (-0.03), la performance énergétique (score DPE : -0.03) ou le nombre d'étages (-0.01), ont un effet négatif plus modeste, cohérent avec une dépréciation liée à la vétusté et à une efficacité énergétique limitée. Enfin, les distances aux entités rouges et vertes apparaissent négligeables dans ce cas précis.

En somme, pour cet individu, les atouts liés à la taille du logement ne compensent pas les effets défavorables de la localisation et de la performance énergétique. La combinaison d'économétrie spatiale et de Machine Learning interprétable offre ainsi un double avantage : elle renforce la robustesse des prédictions et en améliore la lisibilité, ce qui en accroît la pertinence pour des applications concrètes dans l'évaluation ou l'investissement immobilier.

5 Conclusion

Cette recherche a analysé les déterminants des prix immobiliers à Brest, en particulier l'impact du Diagnostic de Performance Énergétique (DPE) dans le cadre de la loi Climat et Résilience. L'approche combinant économétrie spatiale et Machine Learning a permis de mettre en évidence à la fois l'importance des interactions spatiales et le rôle structurant des caractéristiques intrinsèques, environnementales et énergétiques.

Les modèles spatiaux ont confirmé l'existence d'une autocorrélation des prix et ont montré que le DPE constitue un facteur majeur de valorisation ou de décote, avec des effets de débordement sur les biens voisins. En complément, l'utilisation du Random Forest et des Shapley Values a précisé le poids relatif des déterminants (*surface, nombre de pièces, distance au littoral, âge du bâtiment*) tout en offrant une interprétabilité locale des prédictions. Cette capacité à expliquer, bien par bien, l'écart par rapport au prix moyen ouvre la voie à une utilisation opérationnelle des Shapley Values comme outil de scoring, permettant au Crédit Mutuel Arkéa d'évaluer plus finement le positionnement d'un bien sur le marché et d'intégrer ces informations dans ses stratégies de financement et de gestion des risques.