

Machine learning et logements collectifs : quand l'algorithme rencontre la transition énergétique

Elizaveta Logosha*, Myriam Maumy**, Frédéric Bertrand***

@ 84492

Mots-clés : valeur verte, performance énergétique, immobilier durable, transition énergétique, machine learning, science des données

Depuis le début des années 2020, le secteur immobilier français est impacté par la hausse des coûts de construction et les nouvelles réglementations environnementales (RE 2020). Dans ce contexte, la valeur verte, définie comme l'augmentation du prix d'un bien immobilier liée à celle de ses performances énergétiques, devient un enjeu clé. Cet article, réalisé en collaboration avec le Groupe Vivialys, propose une méthode d'estimation de cette valeur à l'aide d'outils issus du machine learning sur des bases de données ouvertes portant exclusivement sur des logements collectifs. Les résultats menés lors de cette collaboration montrent que les logements plus performants énergétiquement bénéficient d'une revalorisation mesurable, offrant ainsi un outil d'aide à la décision pour les acteurs du marché immobilier.

Introduction

Le secteur immobilier français connaît actuellement une transformation majeure, portée par une convergence de défis économiques, environnementaux et réglementaires. La hausse des coûts de construction, alimentée par l'augmentation des prix des matériaux et de la main-d'œuvre sous l'effet de dynamiques économiques et sociétales, s'ajoute à des taux d'intérêt plus élevés, compliquant ainsi l'accès à la propriété. Parallèlement, l'État durcit progressivement les normes en matière de performance énergétique et de réduction de l'empreinte carbone, notamment à travers la RE 2020 [Cabassud, 2024]. Cette nouvelle réglementation impose des exigences accrues aux constructions neuves afin de diminuer leur

consommation énergétique et de favoriser l'utilisation de matériaux plus durables. Dans cette optique, la loi Climat et Résilience¹, adoptée en 2021, vient renforcer les engagements en faveur de la transition écologique du secteur immobilier français. Elle vise notamment une baisse significative des émissions de gaz à effet de serre (GES) et introduit un ensemble de mesures incitatives et coercitives pour accélérer la rénovation énergétique du parc immobilier français existant. Parmi ses principales dispositions figurent le durcissement des critères de performance énergétique des logements et l'interdiction progressive de la mise en location des biens considérés comme des passoires thermiques.

Face à ces transformations, la prise en compte de la valeur verte des logements devient un enjeu majeur pour les acteurs du marché immobilier. En général, la valeur verte d'un bien immobilier désigne l'impact des caractéristiques

* Groupe Vivialys et Université de technologie de Troyes.

** Université de technologie de Troyes et École des hautes études en santé publique.

*** Université de technologie de Troyes.

DÉFINITION PAR NOTAIRES DE FRANCE

La valeur verte définit l'augmentation de valeur engendrée par une meilleure performance énergétique et environnementale d'un bien immobilier par rapport à un autre.

environnementales et énergétiques de ce bien immobilier sur son prix de vente ou de location.

Autrefois considérée comme un critère secondaire, elle occupe désormais une place croissante, tant pour les acquéreurs que pour les promoteurs et les investisseurs. L'amélioration de la performance énergétique des logements, qu'il s'agisse de constructions neuves conformes aux nouvelles réglementations ou de la rénovation du parc immobilier existant, devient ainsi un levier économique et stratégique incontournable pour les acteurs du marché immobilier.

Dans un contexte européen où la transition énergétique guide de plus en plus les décisions des consommateurs, la capacité à quantifier cette valeur verte de manière rigoureuse est essentielle et s'impose comme une priorité. Si de nombreuses études ont mis en évidence un lien entre efficacité énergétique et prix des logements, il reste difficile d'établir des estimations précises et utilisables à l'échelle de l'ensemble du marché. Cette difficulté a amené les auteurs de l'article à proposer des méthodes, basées sur des modèles statistiques, innovantes et visant à estimer de manière robuste la valeur verte, une fois celle-ci définie.

Dans ce travail, les auteurs ont placé le diagnostic de performance énergétique (DPE, voir la Figure 1) au cœur de leur approche pour évaluer la valeur verte des logements. Indispensable pour toute transaction immobilière en France, il attribue aux biens immobiliers une étiquette allant de la lettre A (excellente performance énergétique du logement) à la lettre G (logement très énergivore), sur la base de leur consommation d'énergie et de leur impact en termes d'émissions de CO₂. Ce diagnostic joue un rôle clé dans la perception du marché immobilier, influençant les décisions d'achat de l'acquéreur et la valeur attribuée

aux différents types de logements. Dans cette recherche, les auteurs définissent la valeur verte comme le taux d'augmentation du prix d'un bien immobilier lors de son passage d'une étiquette DPE à une autre étiquette plus performante.

Cette définition renvoie à une valorisation observable sur le marché immobilier, fondée sur la performance environnementale perçue du bien. L'analyse mobilise l'étiquette GES issue du DPE comme indicateur principal, afin de cibler plus spécifiquement la dimension climatique de cette performance. L'approche retenue ici se concentre sur l'empreinte carbone estimée du logement, plutôt que sur la performance énergétique ou environnementale locale. L'approche développée dans cet article s'appuie sur un travail de recherche mené dans le cadre d'une thèse CIFRE, réalisée en partenariat entre l'Université de technologie de Troyes et le Groupe Vivialys, promoteur immobilier alsacien, entre septembre 2021 et décembre 2024. Cette collaboration a été initiée pour répondre à une problématique concrète : le groupe souhaitait procéder à la décarbonation de ses bâtiments, tout en cherchant des arguments commerciaux pour valoriser les bénéfices économiques liés aux performances énergétiques, notamment en termes de prix de revente des biens à moyen et à long terme.

Les travaux de recherche ont été initiés sur les données internes du Groupe Vivialys, puis ont été élargis à des bases de données ouvertes afin de proposer une estimation plus robuste et généralisable de la valeur verte. Les bases de données exploitées comprennent : la base de données nationale des bâtiments² (BDNB), la base de demandes de valeurs foncières³ (DVF) et les statistiques locales de l'INSEE⁴. Ces dernières permettent d'intégrer des variables socio-économiques influençant les prix de l'immobilier.

Machine learning et logements collectifs : quand l'algorithme rencontre la transition énergétique

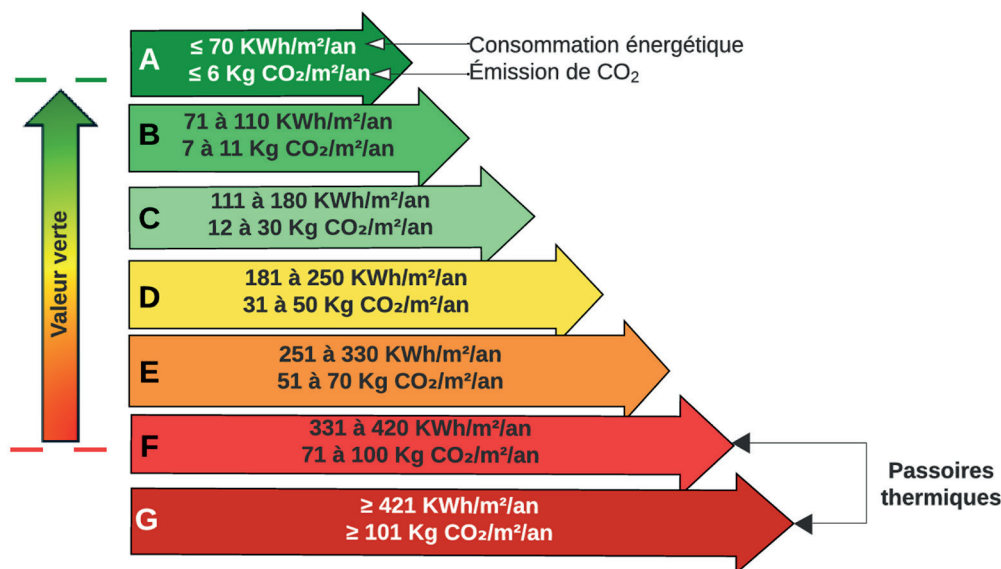


Figure 1. Valeur verte via le diagnostic de performance énergétique

L'objectif de cet article est donc de proposer une estimation de la valeur verte actuelle des logements qui sont sur le marché à l'aide d'outils du *machine learning*, en s'appuyant sur ces bases de données ainsi que sur celles du Groupe Vivialis. À travers des approches comme les forêts aléatoires [Genuer et Poggi, 2019] et l'*eXtreme Gradient Boosting* (XGBoost) [Chen et Guestrin, 2016], nous cherchons à identifier et quantifier l'effet de la transition énergétique sur les prix des logements. Les résultats visent à fournir un cadre d'évaluation qui pourra être utilisé par les professionnels de l'immobilier au niveau national pour mieux intégrer les enjeux énergétiques dans leurs stratégies de valorisation.

1. L'essor de la valeur verte : réglementation, fiscalité et prix du marché

La notion de la valeur verte ne s'est pas imposée naturellement sur le marché immobilier. Elle résulte d'une association de facteurs réglementaires, fiscaux et économiques qui ont progressivement modifié la perception et l'attractivité des logements performants énergétiquement. Si la transition énergétique est aujourd'hui une priorité nationale et européenne, son intégration dans le secteur immobilier repose sur une série de mesures incitatives et contraignantes visant à encourager la rénovation des bâtiments existants et à promouvoir la construction de logements plus durables. Cette notion recouvre de nombreuses dimensions — durabilité des matériaux, modes de chauffage, émissions locales de polluants, empreinte carbone, biodiversité, etc. — mais l'analyse présentée ici se concentre sur la performance énergétique et environnementale telle qu'elle est mesurée par le DPE, et sur ses effets économiques : valorisation immobilière,

possibilité de mise en location, ou incitation à la rénovation.

L'évolution des politiques publiques en matière d'efficacité énergétique a joué un rôle central et fondamental dans la valorisation des logements performants. En France, plusieurs réglementations ont marqué un tournant majeur.

- La réglementation environnementale 2020 (RE 2020) : entrée en vigueur en janvier 2022, elle impose des normes strictes aux constructions immobilières neuves, visant à améliorer leur performance énergétique et à réduire leur empreinte carbone en favorisant des matériaux biosourcés⁵ et des modes de chauffage moins énergivores. Cette réglementation influe directement sur la valeur des biens immobiliers, en rendant les bâtiments conformes plus attractifs pour les acquéreurs et investisseurs, sous réserve que les surcoûts induits restent supportables.
- La loi Climat et Résilience (2021) : elle a introduit une série de mesures visant à éliminer progressivement les passoires thermiques (logements classés avec les étiquettes F et G selon le DPE, voir Figure 1). Dès l'année 2025, ces logements seront interdits à la location annuelle, et à partir de l'année 2034, cette interdiction s'étendra à l'ensemble des biens classés étiquette E ou inférieure. Cette interdiction crée actuellement une pression sur les propriétaires, les incitant soit à rénover leur bien immobilier, soit à vendre à moindre coût, ce qui accentue l'impact du DPE sur la valorisation des logements.
- Le diagnostic de performance énergétique : devenu obligatoire pour toute transaction immobilière, ce diagnostic créé en 2006 classe les bâtiments et les logements de la lettre A, très performants, à la lettre G, très énergivores, selon leur consommation d'énergie et leurs émissions de dioxyde de carbone. Dans ce travail de recherche, nous avons fait du DPE le principal référentiel pour définir et estimer la valeur verte. Une partie des étiquettes utilisées dans l'échantillon provient de

diagnostics réalisés avant la réforme de juillet 2021, date à laquelle les étiquettes énergie et GES, auparavant séparées, ont été fusionnées en une étiquette unique reposant sur la plus mauvaise des deux. Ce choix méthodologique, centré sur l'étiquette GES, distingue cette étude d'autres travaux fondés sur l'étiquette énergétique.

Pour inciter les ménages français à réaliser des travaux de rénovation permettant d'optimiser la performance énergétique de leurs logements, l'État a mis en place, depuis ces dernières années, plusieurs aides financières et dispositifs fiscaux :

- MaPrimeRénov'⁶ : ce programme de subventions couvre une fraction des dépenses engagées pour la rénovation énergétique, avec des montants pouvant atteindre plusieurs milliers d'euros, selon le niveau de revenus du foyer et la nature des travaux effectués.
- Le prêt à taux zéro (PTZ) pour la rénovation : accessible aux ménages sans conditions de ressources, il permet de financer des travaux visant à améliorer l'efficacité énergétique des logements anciens.
- Le dispositif Denormandie : destiné aux investisseurs, il offre des avantages fiscaux aux acheteurs de logements à rénover dans certaines zones définies par l'État.

D'autres dispositifs existent également, tels que la loi Malraux ou les aides à la restauration de monuments historiques. Toutefois, ceux-ci ne sont pas abordés dans cet article, car ils concernent des catégories spécifiques de biens patrimoniaux, distinctes des logements standards étudiés dans le cadre de la valeur verte.

Ces incitations ont pour objectif de réduire le coût des rénovations énergétiques des logements et d'encourager les propriétaires à améliorer leur bien pour éviter une perte de valeur sur le marché immobilier. Toutefois, leur efficacité dépend fortement de la rentabilité des travaux et de la capacité des ménages à les financer.

Machine learning et logements collectifs : quand l'algorithme rencontre la transition énergétique

L'influence de la performance énergétique sur les prix immobiliers est aujourd'hui bien documentée. Plusieurs études⁷ montrent que les logements bien classés selon le DPE bénéficient d'une meilleure valorisation que ceux considérés comme énergivores. Cette tendance s'explique par plusieurs facteurs :

- La perception des acheteurs : un logement performant énergétiquement est perçu comme plus économique à long terme, grâce à des factures de chauffage réduites et un meilleur confort thermique. Ce dernier renvoie à une sensation de bien-être ressentie dans un espace intérieur, résultant d'un équilibre entre la chaleur corporelle et les conditions ambiantes.
- La pression réglementaire : avec les interdictions progressives de location des passoires thermiques, les acheteurs anticipent la nécessité de travaux et ajustent leur prix d'achat en conséquence.
- L'offre et la demande : l'impact du DPE sur le prix d'un bien immobilier varie fortement selon la tension du marché local. Dans les zones où l'offre est abondante, les acheteurs disposent de marges de comparaison plus larges et peuvent intégrer pleinement le critère de performance énergétique dans leur décision. À l'inverse, dans les marchés restreints ou peu concurrentiels, le choix étant plus limité, la qualité énergétique est souvent reléguée au second plan. Par ailleurs, certaines régions proposent essentiellement des biens aux faibles performances énergétiques, faute d'offre rénovée, ce qui contraint les acheteurs à faire des arbitrages. Enfin, dans les secteurs les plus chers, les acquéreurs — souvent plus solvables — peuvent accorder moins d'importance immédiate à l'étiquette DPE, intégrant les coûts de rénovation dans le prix global du projet immobilier.

Toutefois, l'évaluation précise de cet effet est complexe. L'hétérogénéité du marché immobilier, les différences de typologie des biens et les facteurs socio-économiques locaux rendent difficile

l'estimation d'un impact uniforme de la valeur verte. C'est pourquoi nous avons choisi de recourir à des méthodes d'apprentissage automatique (ou *machine learning* en anglais) [Hastie *et al.*, 2009], permettant de modéliser ces variations et d'aboutir à une estimation plus robuste.

2. Des données à la prédiction : les sources utilisées pour estimer la valeur verte

L'estimation de la valeur verte repose sur l'exploitation de diverses bases de données variées, permettant d'analyser l'impact des performances énergétiques sur les prix immobiliers. Dans ce travail exploratoire, nous avons utilisé à la fois des bases de données ouvertes et des données internes du Groupe Vivalys, afin de construire un cadre d'analyse robuste et représentatif du marché immobilier français.

Pour obtenir une analyse généralisable, pérenne et indépendante des spécificités d'un promoteur immobilier en particulier, nous avons mobilisé plusieurs bases de données publiques que nous allons brièvement présenter ci-dessous.

- La base de données nationale des bâtiments (BDNB) : cette base fournit des informations détaillées sur le parc des bâtiments existants en France métropolitaine (résidentiels ou tertiaires), notamment leurs caractéristiques physiques, leur date de construction, leur usage et leur performance énergétique. Construite par croisement géospatial d'une vingtaine de bases de données issues d'organismes publics, elle couvre plus de 32 millions de bâtiments, offrant ainsi un référentiel exhaustif pour l'analyse du parc immobilier existant. L'extraction utilisée dans cette étude date de 2024 et reflète l'état du parc bâti à fin 2023. Elle permet de relier la typologie des bâtiments à leurs performances énergétiques, un élément clé pour la définition et l'estimation de la valeur verte.
- La base de demandes de valeurs foncières (DVF) : mise à disposition par la direction générale des finances publiques (DGFIP),

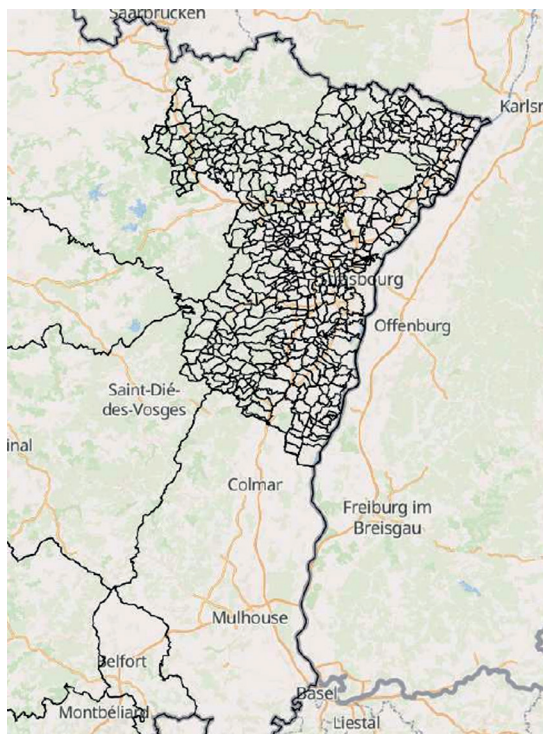


Figure 2. Séparation du département du Bas-Rhin par commune via la base DVF

Source : Etalab, <https://app.dvf.etalab.gouv.fr/>

cette base contient l'ensemble des transactions immobilières enregistrées en France depuis 2014. Elle permet d'accéder à des données précises sur les prix de vente des logements dans toute la France, leur localisation, leur surface, ainsi que d'autres attributs essentiels à l'analyse des déterminants du prix immobilier. Les fichiers DVF, mis à jour semestriellement par la DGFIP, fournissent les transactions immobilières des cinq dernières années. Dans le cadre de cette étude, les données exploitées couvrent donc la période de 2019 à 2024, en fonction des mises à jour disponibles à la date de traitement.

- Les statistiques locales de l'INSEE : ces données permettent d'intégrer des facteurs socio-économiques influençant les prix immobiliers, tels que le revenu médian, la densité de population ou encore l'évolution du marché du

logement à l'échelle communale. Les données utilisées dans cette étude correspondent aux derniers jeux de données disponibles au moment de l'analyse (2024).

- Le diagnostic de performance énergétique : comme mentionné précédemment, nous avons utilisé les étiquettes DPE comme référentiel principal pour définir la valeur verte. Plus précisément, l'analyse repose sur l'étiquette GES, qui classe les logements de A à G selon leurs émissions estimées. Cette approche repose sur l'hypothèse que l'amélioration de cette étiquette génère un effet de valorisation, mesurable via les transactions immobilières.

En complément des bases de données ouvertes, nous avons exploité les données internes du Groupe Vivialys, promoteur immobilier alsacien et partenaire de cette recherche. Celles-ci portent principalement sur des logements construits au cours des cinq années précédant l'étude, offrant ainsi une vision récente des opérations immobilières neuves. Elles ont permis d'obtenir une perspective plus détaillée sur la construction neuve et les stratégies de valorisation des logements économes en énergie. Elles incluent les éléments ci-après :

- Les caractéristiques techniques des logements neufs développés par le groupe, notamment les matériaux utilisés, les équipements énergétiques installés et les choix architecturaux influençant la consommation d'énergie.
- Les prix de vente pratiqués sur différents programmes immobiliers permettent d'observer les écarts de valorisation entre des logements présentant différents niveaux de performance énergétique et environnementale au sein d'un même secteur.
- Les retours clients et tendances commerciales, offrant une vision sur les critères d'achat des clients et leur perception de la valeur verte dans la prise de décision.

Ces données internes ont permis d'affiner le modèle que nous avons construit en validant

Machine learning et logements collectifs : quand l'algorithme rencontre la transition énergétique

certaines tendances observées sur le marché global. En particulier, elles nous ont aidés à mieux comprendre la sensibilité des acheteurs aux critères énergétiques et à tester différentes hypothèses sur l'effet de la performance énergétique sur les prix.

L'utilisation conjointe de données ouvertes et internes constitue un atout majeur pour cette recherche. Tandis que les données publiques permettent une analyse large et généralisable, les données internes apportent une granularité supplémentaire en offrant des informations plus précises sur la construction neuve et le positionnement commercial des biens performants énergétiquement.

Après croisement et nettoyage des données issues des bases BDNB, DVF et INSEE, le jeu final utilisé pour l'analyse regroupe plus de 600 000 transactions immobilières, exclusivement sur des logements collectifs. Ces transactions couvrent la période 2019–2024, conformément à la période disponible dans la base DVF au moment de l'étude.

Le choix de se concentrer sur les appartements s'explique par l'orientation initiale du projet, mené en collaboration avec Trianon Résidences,

marque du Groupe Vivialys spécialisée dans le logement collectif. L'objectif était d'apporter des éléments de réponse concrets à cette typologie de produits. Il est tout à fait envisageable d'étendre cette approche aux maisons individuelles, qui présentent des logiques spécifiques : notamment une plus grande autonomie des propriétaires pour engager des travaux, ce qui pourrait renforcer l'impact perçu de la performance énergétique sur la valeur. Il serait donc pertinent de répliquer la méthodologie sur ce segment pour comparer les résultats, et affiner les analyses sur la valeur verte selon le type de bien.

Cette ampleur confère une solide base statistique aux modèles développés dans la suite de l'étude. À partir de ces données, nous avons développé une approche basée sur des modèles statistiques et d'apprentissage automatique, visant à estimer l'impact du DPE sur la valorisation des logements. Cette approche est détaillée dans la section suivante, qui présente les méthodes utilisées pour modéliser la valeur verte et comparer les performances des différents modèles prédictifs.

Nous avons sélectionné les caractéristiques principales des appartements issues de la BDNB, ainsi que leurs diagnostics de performance énergétique, afin d'assurer la cohérence avec les

Variable	Description	Unité/Format	Source
code_commune	Code INSEE de la commune	Texte	BDNB
prix_m2	Prix par mètre carré issu de la valeur foncière	Euro	DVF
surface_hab	Surface habitable	m ²	DVF
nombre_pièces	Nombre de pièces principales	Nombre décimal	DVF
période_constr	Période de construction du logement	avant 1948, 1948-1974, etc. après 2021	BDNB
année_constr	Année de construction	Année	BDNB
consommation_én	Consommation énergétique	kWh/m ² /an	BDNB
émissions_GES	Émissions de GES	kgCO ₂ /m ² /an	BDNB
étiquette_én	Étiquette énergétique	A, B, C, etc.	BDNB
étiquette_GES	Étiquette GES	A, B, C, etc.	BDNB
type_inst_chauff	Type d'installation de chauffage	Collectif/Individuel	BDNB

Table 1. Variables retenues pour l'estimation de la valeur verte (extrait)

données du Groupe Vivialys. Ces informations ont été fusionnées avec la base DVF, et les logements présentant des données manquantes ont été exclus. La Table 1 présente une partie des variables retenues pour l'analyse.

3. Stratification par commune : une segmentation des marchés locaux via l'algorithme CART

L'immobilier est un marché hétérogène, où les prix et la valorisation des logements varient fortement selon la localisation géographique. Afin de prendre en compte cette diversité, nous avons utilisé les données de l'INSEE pour segmenter les communes françaises en clusters homogènes, en fonction de critères socio-économiques et du marché immobilier local.

Pourquoi une stratification par commune ?

L'impact de la performance énergétique sur la valorisation d'un bien dépend fortement du contexte local. Dans certaines villes où la demande est forte, un bon DPE peut représenter un argument de vente essentiel, tandis que dans d'autres, il peut avoir un effet plus modéré, relégué derrière d'autres facteurs comme l'emplacement ou l'offre disponible. Une stratification par commune permet donc :

- de distinguer les dynamiques de marché : les grandes métropoles, les villes moyennes et les

zones rurales ne réagissent pas de la même manière aux critères énergétiques ;

- d'améliorer la précision des modèles statistiques : en regroupant les communes aux caractéristiques similaires, nous limitons le risque de biais liés aux disparités régionales ;
- de mieux comprendre l'impact du DPE selon les contextes locaux : en comparant les écarts de valorisation entre différents segments du marché.

Pour définir des clusters homogènes de communes, nous avons utilisé l'algorithme CART (*Classification And Regression Trees*) [Breiman *et al.*, 1984], une méthode de segmentation arborescente permettant de regrouper les communes en fonction de plusieurs variables issues des données INSEE, par exemple : densité, taux de logements construits depuis 1990, taux de retraités (plus de 65 ans), nombre de commerces, nombre de services de santé, etc.

L'algorithme CART permet de structurer ces données sous forme d'arbre, divisant les communes en sous-groupes (clusters) ayant des caractéristiques proches. Ainsi, au lieu de considérer la France comme un marché unique, nous avons identifié plusieurs segments de territoires où l'effet de la performance énergétique sur les prix peut être analysé de manière plus fine.

Variable	Description
code_commune	Code postal de la commune
densité	Densité de la population en habitants par kilomètre carré
surface_moyenne	Surface moyenne des logements
prix_moyen_m2	Prix moyen par mètre carré des logements
taux_90_log	Taux de logements construits depuis 1990
taux_enfants	Taux d'enfants (moins de 18 ans)
taux_retrait	Taux de retraités (plus de 65 ans)
services_sante	Nombre de services de santé
sport_culture	Nombre d'établissements sportifs et culturels

Table 2. Variables retenues pour la création des clusters par commune (extrait)

Machine learning et logements collectifs : quand l'algorithme rencontre la transition énergétique

À partir des données communales, sept clusters distincts ont été identifiés par l'algorithme CART, chacun représentant un segment homogène du marché immobilier. Le prix moyen au mètre carré varie significativement d'un cluster à l'autre, allant de 2176 € dans le cluster numéro 1 (zones les moins chères) à 5528 € dans le cluster numéro 7 (zones les plus valorisées) : voir Figure 3. Le prix moyen au mètre carré est présenté ici pour illustrer les différences entre clusters, mais il n'a pas été utilisé comme unique critère de segmentation. Les groupes ont été définis à partir de plusieurs variables socio-économiques communales (revenus médians, densité, part de logements récents, etc.), comme détaillé dans la Table 2.

Cette segmentation reflète essentiellement les dynamiques socio-économiques locales et les niveaux de valorisation immobilière. Toutefois, une autre approche pourrait consister à structurer les territoires en fonction de leur climat, notamment des besoins en chauffage mesurés par le nombre de degrés-jours. Ce critère climatique, non pris en compte dans notre segmentation actuelle, pourrait être intégré dans une future stratification, ou ajouté comme variable explicative dans les modèles, afin d'évaluer plus finement

l'effet du contexte thermique local sur la perception de la performance énergétique.

Afin d'affiner l'estimation de la valeur verte, nous avons développé des modèles distincts adaptés aux différentes catégories de communes identifiées :

- Un modèle spécifique a été développé pour chacun des sept clusters, permettant d'évaluer l'impact du DPE dans des contextes locaux différenciés.
- Un modèle global a également été conçu, intégrant l'ensemble des données pour obtenir une estimation plus large et généralisable.

Cette segmentation assure une meilleure prise en compte des spécificités locales et permet d'éviter les biais liés aux fortes disparités du marché immobilier français.

4. Estimation de la valeur verte : approches statistiques et algorithmiques

L'estimation de la valeur verte repose sur des techniques avancées d'analyse des prix immobiliers afin d'estimer l'impact de la performance énergétique sur la valorisation des biens. Pour atteindre cet objectif, nous avons combiné des approches traditionnelles issues de l'économie immobilière avec des méthodes d'apprentissage automatique, permettant de capturer des relations complexes entre les caractéristiques des logements et leurs prix de transaction.

Modèles hédoniques : fondements et limites pour l'estimation de la valeur verte

L'analyse des prix immobiliers repose traditionnellement sur la modélisation par régression hédonique [Court, 1939], une méthode largement utilisée pour examiner les déterminants du prix des logements. Le terme «hédonique» trouve son origine dans le grec ancien et fait référence au plaisir ou, dans un cadre économique, à l'utilité qu'un acheteur attribue aux caractéristiques d'un bien. Appliquée au marché immobilier,

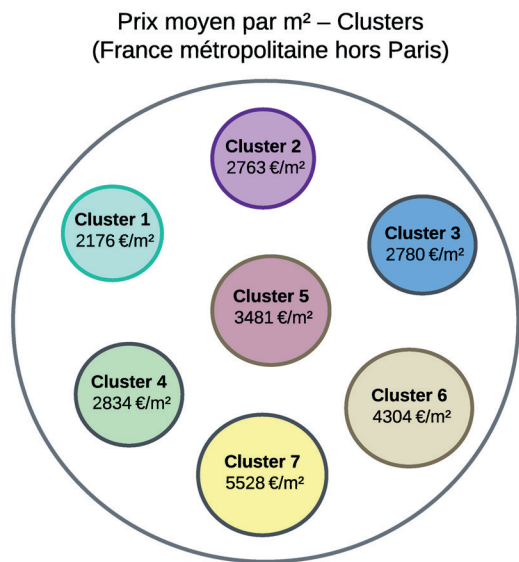


Figure 3. Prix moyen par mètre carré
selon les sept clusters

cette approche permet de décomposer la valeur d'un logement en une somme de ses attributs spécifiques (superficie, localisation, qualité des matériaux, performance énergétique, etc.), en supposant que les différences de prix observées sont en grande partie dues à ces caractéristiques intrinsèques.

Les modèles hédoniques sont couramment employés pour quantifier l'impact des critères environnementaux et énergétiques sur la formation des prix. Plusieurs études ont montré que les certifications environnementales et la performance énergétique influencent significativement la valorisation des biens immobiliers.

Fuerst et McAllister [2011a ; 2011b] utilisent cette approche pour isoler l'effet des certifications environnementales sur les prix des immeubles de bureaux, en prenant en compte des facteurs comme l'âge, la hauteur et l'emplacement des bâtiments. De même, Eichholtz *et al.* [2010 ; 2013] appliquent un modèle similaire pour analyser l'impact des certifications vertes sur le marché américain, tandis que Wiencke [2013] l'a étudié sur le territoire suisse et Ott et Hahn [2018] sur le marché allemand.

Dans ce travail, huit modèles ont été construits afin d'estimer l'effet de la valeur verte sur les prix des logements : un modèle global et un modèle spécifique pour chacun des sept clusters. Nous définissons la valeur verte, au sens environnemental et non uniquement énergétique, comme le taux d'augmentation du prix d'un logement lors du passage d'une étiquette GES à une étiquette plus performante (que nous désignons par la suite comme étiquette environnementale ou climatique, par opposition à l'étiquette énergétique qui concerne la consommation d'énergie). La Table 3 illustre les estimations obtenues avec le modèle hédonique global. Chaque cellule du tableau indique la variation estimée de la valeur d'un logement lors du passage d'une étiquette à une autre. La lecture se fait de la colonne vers la ligne : on compare un logement classé selon l'étiquette de la colonne (étiquette de départ) à un logement identique classé selon l'étiquette de la ligne (étiquette d'arrivée). Par exemple, une cellule contenant «4,16» dans la ligne B et la colonne C signifie qu'un logement classé B vaut en moyenne 4,16 % de plus qu'un logement identique classé C. Le tableau n'est pas symétrique, ce qui reflète la nature du modèle utilisé pour ces estimations. Les valeurs négatives sur la diagonale supérieure, qui représentent des incohérences

de vers	A	B	C	D	E	FG
A	0,00	-1,53	2,57	6,09	4,48	3,06
B	1,55	0,00	4,16	7,73	6,10	4,65
C	-2,51	-4,00	0,00	3,43	1,85	0,47
D	-5,74	-7,18	-3,31	0,00	-1,52	-2,86
E	-4,29	-5,75	-1,82	1,54	0,00	-1,36
FG	-2,97	-4,45	-0,47	2,94	1,38	0,00

Table 3. Valeur verte via modèle hédonique : modèle global

Machine learning et logements collectifs : quand l'algorithme rencontre la transition énergétique

dans les relations attendues entre les étiquettes, sont mises en évidence en rouge dans la table des résultats.

Ces incohérences résultent des limites structurelles des modèles hédoniques, notamment de leur difficulté à représenter des effets non linéaires ou des interactions entre caractéristiques. Elles peuvent également refléter des dimensions non modélisées, comme l'esthétique architecturale (parfois critiquée dans les bâtiments très performants) ou la perception subjective du confort. Malgré leur efficacité, les modèles hédoniques présentent certaines limites. Ils supposent une relation linéaire entre les caractéristiques du logement et son prix, ce qui peut être trop simpliste pour capturer l'ensemble des dynamiques du marché. De plus, ils sont sensibles aux variables omises et ne permettent pas toujours de saisir les interactions complexes entre les différents facteurs influençant la valorisation des logements.

Face à ces limites, nous avons exploré des approches plus avancées. Nos résultats suggèrent que nous pourrions construire des modèles plus performants, mieux adaptés aux spécificités du marché immobilier et de la valeur verte. Pour cela, nous nous tournons vers le *machine learning*, une approche plus flexible et capable de capturer des relations non linéaires entre les variables.

5. Modèles non linéaires pour l'estimation de la valeur verte : forêts aléatoires et XGBoost

Afin de surmonter les limites des modèles hédoniques, nous avons eu recours à des méthodes d'apprentissage automatique (ou *machine learning*), couramment utilisées dans d'autres domaines comme la santé, l'énergie ou la météorologie [Hastie *et al.*, 2009]. Ces approches permettent de modéliser des relations complexes entre variables sans faire d'hypothèses rigides sur leur forme.

Parmi ces méthodes, les forêts aléatoires fonctionnent en construisant un ensemble d'arbres de décision à partir de sous-échantillons aléatoires des données. Chaque arbre produit une

prédiction indépendante, et le résultat final est obtenu par agrégation (par exemple, vote majoritaire ou moyenne), ce qui renforce la robustesse du modèle et limite le risque de surapprentissage (*overfitting*) [Breiman, 2001].

XGBoost (*eXtreme Gradient Boosting*) est un algorithme d'agrégation séquentielle fondé sur le principe du *gradient boosting*. Il corrige progressivement les erreurs commises par les modèles précédents et s'illustre par une grande efficacité prédictive ainsi qu'une capacité à gérer des données complexes, bruitées ou présentant des interactions non linéaires [Chen et Guestrin, 2016].

Ces deux techniques sont utilisées ici pour estimer l'effet du DPE sur les prix de l'immobilier et produire des prédictions plus fiables dans un marché hétérogène.

Forêts aléatoires : capturer les interactions complexes

Les forêts aléatoires permettent de modéliser des interactions complexes entre les caractéristiques des logements (étiquette DPE, surface, localisation, etc.) sans imposer de relation linéaire. Elles sont particulièrement adaptées à des jeux de données hétérogènes et segmentés comme le marché immobilier.

Dans notre étude, nous les mobilisons pour estimer l'effet du DPE sur les prix, en tenant compte des dynamiques locales révélées par la stratification en clusters. Cette approche permet ainsi de détecter les différences de valorisation selon les profils territoriaux et la typologie des logements.

XGBoost : une optimisation du *gradient boosting*

XGBoost est un algorithme d'agrégation séquentielle qui construit ses arbres de décision de manière itérative, en corrigeant les erreurs des prédictions précédentes. Il est reconnu pour sa précision, sa gestion efficace des données déséquilibrées et sa capacité à traiter des interactions complexes entre variables.

Dans notre étude, il est utilisé en parallèle des forêts aléatoires pour estimer l'effet du DPE sur les prix immobiliers, à la fois dans un modèle global et dans des modèles différenciés par cluster. Sa flexibilité algorithmique permet d'évaluer finement les contributions des variables énergétiques à la valeur verte selon les segments de marché.

6. Estimation de la valeur verte selon les trois méthodes

Dans ce travail, huit modèles distincts ont été construits pour chacune des trois approches retenues : un modèle global et un modèle pour chacun des sept clusters territoriaux, identifiés via l'algorithme CART. La valeur verte est définie ici comme le taux d'augmentation du prix d'un logement lors du passage d'une étiquette climatique à une étiquette plus performante.

Les modèles hédoniques produisent des valeurs vertes particulièrement élevées. Par exemple, dans les clusters 1 à 3, les gains associés à une amélioration énergétique allant des étiquettes F ou G vers B atteignent jusqu'à 13,21 % pour les clusters correspondant aux segments de marché les moins chers, avec un prix moyen au mètre carré compris entre 2176 € et 2780 €. Le modèle global, quant à lui, indique, par exemple, une augmentation de 7,73 % pour le passage de l'étiquette climatique D à B et 2,57 % pour le passage de l'étiquette climatique C à A.

Toutefois, ces résultats soulèvent des questions quant à leur fiabilité. Les valeurs estimées sont parfois négatives ou incohérentes d'un segment à l'autre, ce qui suggère que cette méthode surestime certains effets et en sous-estime d'autres, en raison de sa rigidité structurelle.

Les modèles hédoniques, largement utilisés dans la littérature, apparaissent ici trop simplificateurs pour capturer les dynamiques complexes du marché immobilier et les effets différenciés des étiquettes environnementales. Ces limites justifient le recours à des approches plus flexibles, capables de mieux représenter les interactions entre variables et de produire des estimations plus robustes.

Forêts aléatoires et XGBoost : vers des estimations plus cohérentes

Pour surmonter ces limites, nous avons mobilisé deux méthodes d'apprentissage automatique : les forêts aléatoires et XGBoost, en y intégrant une analyse des contributions par les valeurs de Shapley [Shapley, 1953]. Cette technique permet de quantifier, de façon équitable, l'effet marginal des étiquettes environnementales sur le prix d'un logement.

La procédure d'estimation de la valeur verte repose sur l'algorithme suivant :

1. Construction du modèle prédictif (forêts aléatoires ou XGBoost) à partir des données sélectionnées.
2. Approximation des valeurs de Shapley par la méthode FastSHAP [Greenwell, 2024], utilisant un algorithme de Monte-Carlo [Štrumbelj et Kononenko, 2014] pour estimer la contribution de chaque variable. Ces valeurs s'inscrivent dans le cadre de SHAP (*SHapley Additive exPlanations*), une méthode interprétative inspirée de la théorie des jeux [Lundberg et Lee, 2017].
3. Extraction des valeurs de Shapley spécifiques aux étiquettes climatiques.
4. Calcul de la valeur verte comme différence moyenne ou médiane des contributions associées à deux étiquettes successives (par exemple, de C vers B).

Ce processus permet de produire des estimations plus stables et mieux adaptées à la complexité des données immobilières que les approches linéaires traditionnelles. Les résultats montrent des niveaux de valorisation comparables, mais globalement plus modérés, à ceux des modèles hédoniques. Pour les forêts aléatoires, la valeur verte moyenne du modèle global est proche de 1 %, ce qui témoigne d'estimations plus prudentes. Pour XGBoost, les estimations sont légèrement plus élevées, allant jusqu'à 3,12 % pour le modèle global.

Machine learning et logements collectifs : quand l'algorithme rencontre la transition énergétique

La Table 4 présente un exemple de matrice d'estimation des valeurs vertes obtenues via les forêts aléatoires sur le modèle global. Chaque cellule indique le gain estimé en pourcentage associé à une amélioration de l'étiquette climatique du logement, selon les moyennes des valeurs de Shapley calculées.

Ce résultat, qui suggère, par exemple, un surcroît de valorisation de 1,35 % pour un passage d'une étiquette D à une étiquette A, peut sembler modeste au regard des économies d'énergie théoriques associées à une telle amélioration. Plusieurs éléments permettent toutefois d'interpréter ce niveau de valorisation. La valeur verte captée ici est fondée sur le consentement à payer observé sur le marché : elle intègre donc les préférences effectives des acquéreurs, qui peuvent sous-valoriser certains gains énergétiques jugés peu visibles ou incertains. Par ailleurs, les effets estimés sont marginaux et conditionnels, mesurés toutes choses égales par ailleurs, dans un modèle contrôlant pour de nombreuses variables explicatives. Enfin, les effets cumulés de plusieurs sauts d'étiquette peuvent être mieux appréhendés en analysant la structure globale de la matrice.

Afin de mieux comprendre l'évolution de la valeur verte selon le contexte local et le cycle de vie des logements, l'analyse a été conduite à la fois par cluster territorial et par période de construction des logements (voir la Table 1), afin d'estimer l'impact des étiquettes climatiques en fonction de l'âge des biens immobiliers. Bien que l'absence de données de transaction dans les départements d'Alsace-Moselle — en raison de leur régime foncier spécifique hérité du droit local — ait limité l'exploitation directe des ventes sur ce territoire, il a été possible, grâce à la classification CART, de repérer les clusters correspondant aux communes où Trianon Résidences, une des marques du Groupe Vivialys, développe l'essentiel de ses programmes. Ces communes se trouvent majoritairement dans les clusters 4 et 6, regroupant des villes comme Strasbourg, Mulhouse, Brumath ou Illkirch-Graffenstaden.

Les résultats de notre estimation montrent que la valeur verte augmente avec l'ancienneté du bien, les logements récents affichant des gains plus modestes, les plus anciens des hausses plus marquées. Ces gains correspondent à l'effet estimé d'un passage d'une étiquette GES à l'étiquette immédiatement supérieure. Par exemple,

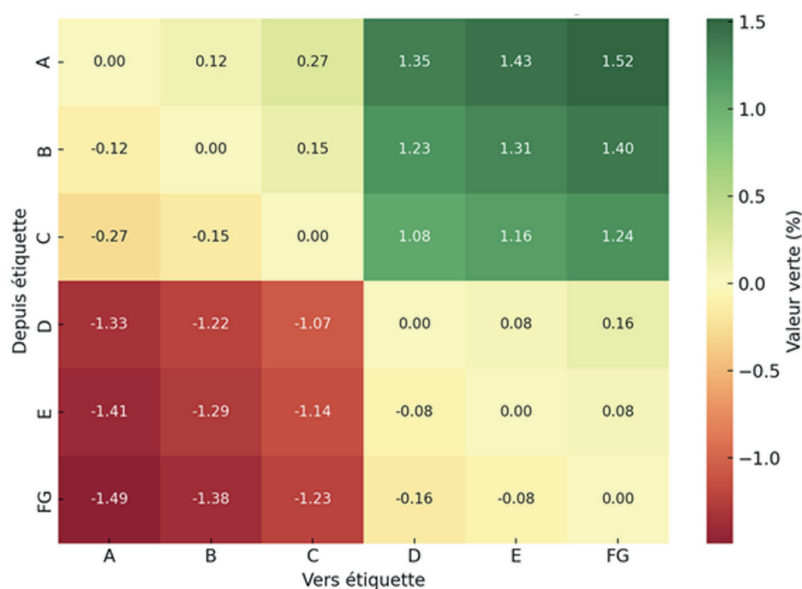


Table 4. Valeur verte via forêts aléatoires : modèle global

dans les clusters 4 et 6, les logements construits depuis une dizaine d'années affichent une valorisation verte allant de 1 % à 1,64 %, tandis que ceux construits il y a une trentaine d'années présentent des hausses pouvant atteindre près de 4 %. Le modèle global, de son côté, indique des hausses de 1 %, 1,77 % et 2,46 % pour les logements de 10, 20 et 30 ans respectivement.

En se basant sur ces éléments, ainsi que sur les rapports des Notaires de France, nos propres estimations par cluster et période de construction suggèrent qu'à l'horizon de 10 ans, les logements conformes aux normes environnementales les plus exigeantes pourraient se vendre environ 3 % plus cher, et jusqu'à 6 % à l'horizon de 20 ans.

Ces projections, bien qu'indicatives, confirment la dynamique croissante autour de la valeur verte et l'intérêt d'outils prédictifs robustes pour en anticiper les effets économiques.

Ces résultats soulignent l'importance croissante accordée aux performances environnementales dans la valorisation des biens, notamment pour les logements plus anciens, où l'amélioration de l'étiquette climatique est perçue comme un signal fort de qualité et d'efficacité énergétique.

Comparaison de la qualité d'ajustement des modèles

L'évaluation des performances prédictives des modèles repose principalement sur le coefficient de détermination R^2 , indicateur de la part de variance expliquée par les modèles. Ces résultats permettent d'apprécier non seulement l'efficacité des approches statistiques et algorithmiques, mais aussi la robustesse des estimations produites.

Les modèles hédoniques, bien que couramment utilisés dans la littérature immobilière, présentent dans notre cas des limites importantes en matière de qualité d'ajustement. Le modèle global affiche un R^2 de 0,474, et les résultats par cluster varient fortement, avec des valeurs de R^2 allant jusqu'à 0,620, mais également des performances bien moindres dans plusieurs groupes. Ce manque de stabilité suggère une incapacité à

capturer les effets non linéaires et les interactions complexes entre les variables. En outre, l'apparition de valeurs incohérentes, voire négatives, dans l'estimation de la valeur verte révèle un risque de surestimation ou sous-estimation systématique de certaines étiquettes, probablement lié à la structure rigide du modèle.

Il est également probable que l'absence de variables de localisation plus précises — au-delà du simple code postal ou cluster — limite la capacité explicative du modèle, en ne tenant pas compte des dynamiques microterritoriales influençant les prix immobiliers.

À l'inverse, les méthodes d'apprentissage automatique offrent des performances nettement supérieures. Le modèle global fondé sur les forêts aléatoires atteint un R^2 de 0,793, tandis que celui construit avec XGBoost affiche un R^2 de 0,786. Cette amélioration significative démontre la capacité de ces modèles à intégrer les relations non linéaires entre les caractéristiques des logements, à gérer les interactions multiples, ainsi qu'à mieux exploiter la diversité des données.

Les résultats par cluster confirment également cette tendance : les algorithmes de *machine learning* fournissent des modèles plus stables et plus précis, y compris dans les segments de marché où les modèles linéaires échouaient à produire des estimations cohérentes. De manière générale, les forêts aléatoires montrent une légère supériorité en termes de précision, tandis que XGBoost offre une meilleure flexibilité sur les ensembles de données complexes.

Ainsi, du point de vue de la qualité d'ajustement, les approches d'apprentissage automatique apparaissent comme particulièrement bien adaptées pour estimer ce que l'on appelle la « valeur verte faible », c'est-à-dire la variation marginale du prix d'un bien immobilier associée à un changement de performance énergétique, toutes choses égales par ailleurs. Contrairement aux méthodes économétriques classiques, qui imposent des formes fonctionnelles *a priori* et peuvent difficilement capter des relations non linéaires complexes ou des interactions multiples entre variables,

Machine learning et logements collectifs : quand l'algorithme rencontre la transition énergétique

les méthodes d'apprentissage automatique (comme les forêts aléatoires ou XGBoost) sont capables de modéliser de manière souple des relations hétérogènes et localement spécifiques. Cette capacité à s'adapter aux structures de données, même en présence de bruit, de multicollinéarité ou de déséquilibre dans les classes d'étiquette énergétique, en fait des outils robustes dans un marché immobilier marqué par une forte hétérogénéité spatiale et structurelle [Mullainathan et Spiess, 2017; Athey, 2019]. De plus, leur performance prédictive est souvent supérieure à celle des modèles linéaires, ce qui en fait des outils particulièrement pertinents dans un contexte de mutation du parc immobilier où les effets liés à la transition énergétique sont encore diffus et difficiles à isoler par des approches standards.

7. De l'analyse à l'action : la suite logique

Bilan des résultats et enjeux pour les acteurs du marché

Les travaux présentés ont permis de proposer une méthode rigoureuse d'estimation de la valeur verte, fondée sur l'analyse de grandes bases de données immobilières et sur l'application de modèles de *machine learning* interprétables. Les résultats confirment que les logements présentant de meilleures performances climatiques bénéficient, en moyenne, d'une valorisation positive à la revente, avec des effets plus marqués pour les biens anciens. Par exemple, dans certains clusters, un logement datant de 30 ans peut voir sa valeur augmenter de près de 4 % lorsqu'il dispose d'une étiquette énergétique plus performante. Le modèle global, quant à lui, indique des hausses allant de 1 % à 2,5 % selon l'âge du bien. Ces résultats rejoignent, tout en les nuanciant, les constats des Notaires de France, qui rapportent des hausses pouvant atteindre 20 % dans certaines régions. Cette différence s'explique en grande partie par le fait que ces travaux s'appuient sur l'étiquette énergétique du DPE, plus visible pour les acheteurs, alors que notre étude se concentre sur l'étiquette GES, reflétant l'empreinte carbone du logement. Les comportements d'achat étant souvent orientés par la perspective

de dépenses futures plutôt que par des critères environnementaux, la valorisation de l'étiquette GES peut apparaître moins marquée. Par ailleurs, les modèles hédoniques linéaires utilisés dans de nombreuses études antérieures tendent à surestimer l'effet des étiquettes environnementales, parfois de manière significative.

Dans certaines simulations réalisées à titre comparatif, l'effet estimé atteignait 13 à 15 % dans les clusters 2 et 3, ce qui paraît excessif au regard des dynamiques réellement observées sur le marché. À l'inverse, les modèles non paramétriques mobilisés ici peuvent conduire à une légère sous-estimation, notamment en atténuant les effets marginaux extrêmes.

Par ailleurs, l'analyse repose sur l'étiquette GES et non sur la consommation énergétique. Ce choix, cohérent avec l'objectif d'évaluer la dimension environnementale des logements, ne correspond pas nécessairement aux critères les plus visibles ou déterminants pour les acquéreurs, souvent plus sensibles aux coûts futurs qu'à l'empreinte carbone. Cette différence de référentiel contribue à expliquer l'écart avec les études s'appuyant sur le DPE énergétique. De plus, une partie des logements inclus dans le périmètre de l'étude a été évaluée à partir d'anciennes versions du DPE, dans lesquelles les étiquettes énergétique et GES étaient dissociées. Cette hétérogénéité dans les référentiels d'évaluation peut avoir réduit la cohérence de la variable d'analyse et atténué l'ampleur des effets mesurés.

Ainsi, si les résultats obtenus apparaissent modestes en valeur absolue, ils correspondent à une estimation prudente, ancrée dans les comportements effectifs observés sur le marché et dans les données disponibles.

Néanmoins, ces estimations restent conditionnées à la stabilité des marchés de l'énergie et des matériaux. Toute évolution réglementaire ou économique majeure pourrait venir renforcer, ou au contraire atténuer, l'effet de la performance énergétique sur la valeur immobilière. Ces estimations reflètent ainsi le consentement à payer observable à un instant donné, sans intégrer

explicitement les anticipations sur l'évolution des prix de l'énergie. Or, si ces prix poursuivent leur hausse dans les prochaines décennies, la valorisation d'un logement économe pourrait mécaniquement s'accroître. La performance énergétique apparaît alors comme un levier stratégique de long terme, justifiant de futures analyses intégrant une dynamique temporelle plus fine. En parallèle, ces estimations doivent également être interprétées à la lumière des contraintes d'accès au capital des ménages. La capacité à financer un logement plus performant énergétiquement ne dépend pas uniquement d'une préférence environnementale, mais aussi de conditions de financement (revenus, épargne, solvabilité, etc.). Si la stratification par clusters permet de différencier des segments de marché sur des bases socio-économiques et de prix, elle ne permet pas de capter finement cette variable essentielle. Une extension de ce travail pourrait intégrer des indicateurs d'accès au crédit ou de solvabilité locale, afin d'isoler plus précisément la part de consentement à payer librement exprimée.

Pour les promoteurs, cette valorisation peut justifier des choix de conception plus ambitieux sur le plan environnemental, en particulier dans la construction neuve. Pour les investisseurs, elle offre une meilleure lisibilité sur le potentiel de valorisation à moyen terme, notamment dans les territoires où les exigences climatiques se renforcent. Enfin, pour les acquéreurs, la valeur verte représente à la fois un signal environnemental et, indirectement, un facteur de sécurité économique face à la volatilité énergétique et aux futures restrictions réglementaires. En somme, la reconnaissance croissante de la valeur verte dans les décisions de marché souligne la nécessité d'outils fiables pour l'identifier, l'estimer et la valoriser.

Perspectives et développement futur de la valeur verte

Les travaux présentés confirment que la valeur verte constitue un levier stratégique croissant pour les acteurs de l'immobilier, en particulier dans un contexte de transition énergétique accélérée. Les méthodes mobilisées — des modèles hédoniques aux approches de *machine learning*

interprétées par les valeurs de Shapley — ont permis de quantifier de manière rigoureuse l'effet des performances environnementales sur la valorisation des logements.

Il aurait été pertinent d'étudier l'évolution temporelle de la valeur verte sur une période plus longue, notamment pour identifier d'éventuelles tendances liées à l'impact des politiques publiques ou à la sensibilité croissante des acheteurs. Toutefois, au moment de cette recherche, les données DVF accessibles ne couvraient qu'une période limitée (2019–2024). Cette analyse pourrait faire l'objet d'un prolongement futur à mesure que de nouvelles données seront publiées.

Ce travail s'appuie sur les étiquettes DPE, aujourd'hui largement utilisées dans les transactions immobilières et dans la plupart des travaux sur la valeur verte. Bien que leur mode de calcul fasse l'objet de débats, elles constituent à ce jour le seul indicateur environnemental disponible à l'échelle nationale et intégrable dans des bases de données ouvertes. La méthode proposée ici, fondée sur des modèles interprétables de *machine learning*, reste entièrement généralisable et pourrait être mobilisée à l'avenir avec d'autres indicateurs, si de nouveaux outils plus représentatifs de la performance environnementale venaient à être déployés. Cette ouverture méthodologique s'inscrit dans une logique d'adaptation continue aux évolutions du secteur, tant réglementaires que techniques. Ce chantier s'inscrit pleinement dans les priorités du Groupe Vivalys, en lien avec ses engagements en matière de décarbonation du bâti, et ouvre la voie à une nouvelle génération d'indicateurs environnementaux immobiliers, articulés autour de l'innovation, de la durabilité et de la valorisation économique des pratiques vertueuses. En ce sens, la valeur verte n'est pas seulement une variable économique : elle devient un vecteur stratégique de transformation du secteur immobilier.

Machine learning et logements collectifs : quand l'algorithme rencontre la transition énergétique

REMERCIEMENTS

Ce travail a été réalisé en collaboration avec le Groupe Vivialys, que nous remercions pour son implication tout au long de cette recherche, ainsi que son équipe de direction pour la qualité des échanges. Nous adressons une mention particulière à Cédric Simonin, encadrant industriel de la thèse à l'origine de cet article, pour son accompagnement et ses contributions tout au long du projet.

NOTES

1. La loi n° 2021-1104 du 22 août 2021, portant lutte contre le dérèglement climatique et renforcement de la résilience face à ses effets.
2. Centre scientifique et technique du bâtiment, <https://www.data.gouv.fr/fr/datasets/base-de-donnees-nationale-des-batiments/>.
3. Ministère de l'Économie, des Finances et de l'Industrie, <https://www.data.gouv.fr/fr/datasets/demandes-de-valeurs-foncières/>.
4. Institut national de la statistique et des études économiques, <https://statistiques-locales.insee.fr/>.
5. Les matériaux biosourcés sont des matériaux provenant de la matière organique renouvelable, d'origine végétale ou animale, et peuvent servir de matière première pour des produits de construction, de décoration, de mobilier fixe, ainsi que pour des éléments de construction dans les bâtiments.
6. MaPrimeRénov', <https://www.maprimerenov.gouv.fr/>.
7. Notaires de France, 2021, 2022, 2023. «La valeur verte des logements», Rapports d'étude annuels.

RÉFÉRENCES

- Athey S., 2019. "The Impact of Machine Learning on Economics", *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press, pp. 507-547.
- Breiman L., Friedman J.H., Olshen R.A., Stone C.J., 1984. *Classification and Regression Trees*, Wadsworth, Belmont.
- Breiman L., 2001. "Random Forests", *Machine Learning*, 45(1), 5-32.
- Cabassud N., 2024. «Guide RE 2020. Réglementation environnementale», ministère de la Transition écologique et Centre d'études et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement (Cerema).
- Chen T., Guestrin C., 2016. "XGBoost: A Scalable Tree Boosting System", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pp. 785-794, New York, NY, USA, ACM.
- Court A.T., 1939. "Hedonic price indexes with automotive examples", in *The Dynamics of Automobile Demand*, pp. 98-119, General Motors, New York.
- Eichholtz P., Kok N., Quigley J.M., 2010. "Doing well by doing good? Green office buildings", *American Economic Review*, 100(5):2492-2509.
- Eichholtz P., Kok N., Quigley J.M., 2013. "The economics of green building", *Review of Economics and Statistics*, 95(1):50-63.
- Fuerst F., McAllister P., 2011a. "Eco-labeling in commercial office markets: Do LEED and Energy Star offices obtain multiple premiums?", *Ecological Economics*, 70(6):1220-1230.
- Fuerst F., McAllister P., 2011b. "Green Noise or Green Value? Measuring the Effects of Environmental Certification on Office Values", *Real Estate Economics*, 39(1):45-69.
- Genuer R., Poggi J.-M., 2019. *Les forêts aléatoires avec R*, Presses Universitaires de Rennes, coll. Pratique de la statistique.

Greenwell B.M., 2024. "fastshap: Fast Approximate Shapley Values", CRAN – Comprehensive R Archive Network.

Hastie T., Tibshirani R., Friedman J., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2^e édition, Springer.

Lundberg S.M., Lee S.-I., 2017. "A Unified Approach to Interpreting Model Predictions", *Advances in Neural Information Processing Systems*, vol. 30.

Mullainathan S., Spiess J., 2017. "Machine Learning: An Applied Econometric Approach", *Journal of Economic Perspectives*, 31(2), pp. 87-106.

Ott C., Hahn J., 2018. "Green pay off in commercial real estate in Germany: assessing the role of

Super Trophy status", *Journal of Property Investment and Finance*, 36(1):104-124.

Shapley L.S., 1953. "A Value for n-Person Games", Contributions to the Theory of Games, vol. II, Princeton University Press, Princeton, pp. 307-317.

Štrumbelj E., Kononenko I., 2014. "Explaining prediction models and individual predictions with feature contributions", *Knowledge and Information Systems*, vol. 41, pp. 647-665.

Wiencke A., 2013. "Willingness to Pay for Green Buildings – Empirical Evidence from Switzerland", *SSRN Electronic Journal*, 5:1-31.

BIOGRAPHIES

ELIZAVETA LOGOSHA est docteure en sciences pour l'ingénieur. Formée entre Saint-Pétersbourg, Strasbourg et Troyes, elle développe une approche interdisciplinaire mêlant modélisation stochastique et enjeux de durabilité. Dans le cadre de sa thèse à l'Université de technologie de Troyes, elle a travaillé sur l'application du *machine learning* à l'analyse de données issues de la transition énergétique.

MYRIAM MAUMY, docteur en statistique de l'Université Pierre et Marie Curie (Paris VI) en 2002 et titulaire d'une HDR de l'Université de Strasbourg, a commencé sa carrière d'enseignant-chercheur à l'Université de Rennes 2 puis à l'Université de Strasbourg où elle restée 16 ans à développer des relations universités-entreprises. Actuellement, elle occupe un poste de professeur des universités en santé numérique à l'École des hautes études en santé publique.

Ancien élève de l'École normale supérieure de Lyon, titulaire d'un doctorat de l'Université Louis Pasteur et d'une HDR en mathématiques appliquées de l'Université de Strasbourg, **FRÉDÉRIC BERTRAND** est professeur des universités à l'Université de technologie de Troyes où il est responsable de la formation de mastère spécialisé Expert Big Data Engineer et du pilotage, au sein de l'UTT, du projet CMA en IA MacMia et du cluster IA Hi!Paris. Il a travaillé sur la modélisation statistique et l'apprentissage automatique, en particulier avec le langage R auquel il a contribué pour près de vingt packages.