Simulation du gisement de rénovation énergétique et carbone du parc résidentiel français avec la BDNB

Pascal Schetelat[1,*], Antoine Breitwiller[1]

[1]CSTB
24 rue Joseph Fourier - 38400 Saint Martin D'Hères
*pascal.schetelat@cstb.fr

RÉSUMÉ. La simulation énergétique du parc bâti à l'échelle urbaine repose sur des jeux de données incomplets et hétérogènes. En France, les diagnostics de performance énergétique (DPE) couvrent 17% des bâtiments et présentent des biais de sélection. Cet article propose une approche basée sur la donnée visant à reconstruire, de manière probabiliste, les performances des bâtiments à partir de la Base de Données Nationale des Bâtiments (BDNB). La méthode s'appuie sur un modèle génératif conditionnel combinant un auto-encodeur variationnel et des mécanismes d'attention. Pour chaque bâtiment, une population Monte-Carlo est générée puis simulée dans un moteur conforme à la méthode 3CL-DPE 2021. Un cas d'usage évaluant les gisements d'économie d'énergie après rénovation globale est présenté. Ce cadre probabiliste constitue un jumeau énergétique urbain adapté à l'aide à la décision publique ainsi qu'à l'information des particuliers via l'application gorenove

MOTS-CLÉS. monte-carlo, machine learning, simulation urbaine, BDNB, DPE

ABSTRACT. Urban building energy simulation relies on incomplete and heterogeneous datasets. In France, Energy Performance Certificates (EPCs) cover only a fraction of the residential stock and exhibit marked selection biases. This paper proposes a data-driven approach to probabilistically reconstruct building energy performances using the National Building Database (BDNB). The method relies on a conditional generative model combining a variational auto-encoder and attention mechanisms. For each building, a Monte-Carlo population is generated and simulated in an engine complying with the 3CL-DPE 2021 regulatory method. A use case evaluating energy savings potentials after a global renovation is presented. This probabilistic framework serves as an urban energy twin tailored for public decision support.

KEYWORDS. monte-carlo, machine learning, urban simulation, building database, EPC

1. INTRODUCTION

La massification de la rénovation énergétique du parc résidentiel exige une cartographie de référence des performances du bâti qui soit à la fois à jour, exhaustive et cohérente. Ce « jumeau énergétique » est indispensable pour planifier, évaluer les politiques publiques et éclairer la prise de décision de tous les acteurs de la transition, des pouvoirs publics aux particuliers.

Cependant, la simulation énergétique urbaine se heurte classiquement à la rareté des données. En France, seuls 17% des bâtiments disposent d'au moins un DPE récent. De plus, l'actualisation des DPE est liée aux mutations immobilières et non aux travaux, et ils présentent des biais de sélection inhérents aux dispositifs réglementaires. L'incitation à obtenir une meilleure étiquette génère des effets de seuil marqués, et la dépendance de la méthode DPE aux priorités politiques (ex : abaissement du coefficient de conversion de l'énergie primaire électrique à 1,9 (Ministère de la Transition Écologique, 2025)) rend l'approche par classification supervisée en

« boîte noire » fragile. Au-delà de la photographie de l'existant, ce jumeau doit pouvoir rendre compte de l'efficacité des stratégies de rénovation en permettant de simuler des variantes, de réaliser de la prospective ou à rendre compte de phénomènes qui ne seraient pas pris en compte explicitement dans les DPE, comme la surchauffe des bâtiments lors de vagues de chaleur (Kyriakodis, 2025; Merlet et al., 2025).

Pour s'affranchir de ces limites, nous proposons une approche de simulation reposant sur la reconstruction des données d'entrée du calcul DPE méthode 3CL (arrêté 2021). Les géométries et masques solaires sont extraits de la Base de Données Nationale des Bâtiments (BDNB). Les performances manquantes de l'enveloppe et des systèmes sont générées par une approche d'imputation via un encodeur variationnel conditionnel (cVAE). En phase d'inférence, ce modèle génère pour chaque bâtiment un échantillon Monte-Carlo de données probables, préservant la structure de dépendance entre les variables. Propagées à travers la méthode 3CL, ces données permettent de simuler les performances énergétiques probables tout en quantifiant l'incertitude épistémique, fondement indispensable d'un jumeau urbain d'aide à la décision publique.

2. MÉTHODOLOGIE DE RECONSTRUCTION DES DONNÉES MANQUANTES

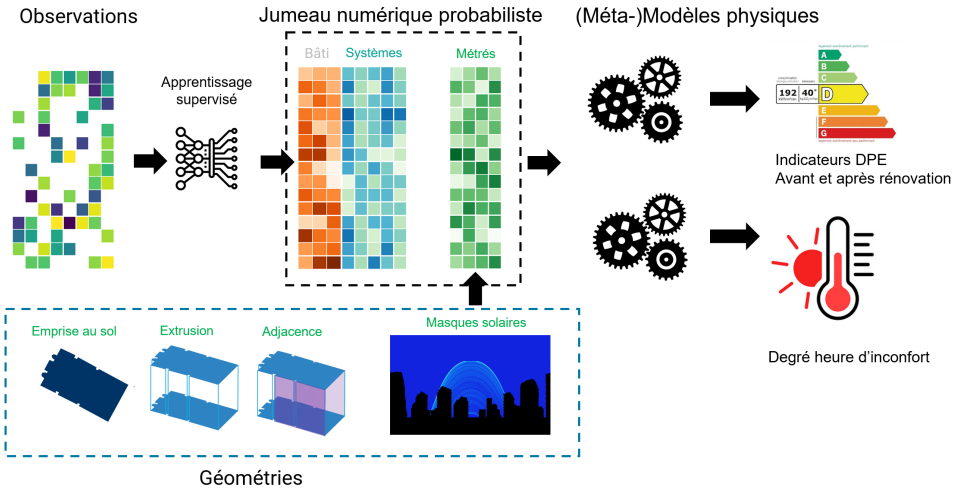
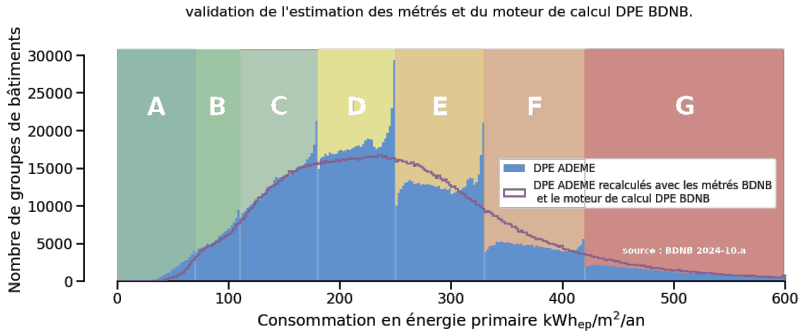


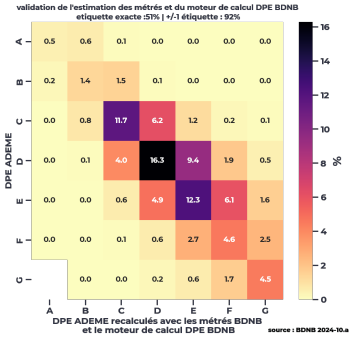
Figure 1 : Schéma de la chaîne de prédiction

2.1. EXPLOITATION DES GÉOMÉTRIES DE LA BDNB

Les surfaces déperditives, les ponts thermiques et les masques solaires de chaque bâtiment sont estimés à partir des emprises au sol extrudées (modélisation 2.5D) en prenant en compte l'adjacence entre bâtiments. Cette approche garantit une prise en compte exhaustive des mitoyennetés et du contexte urbain environnant (masques solaires calculés à chaque face). Cette approche est validée en re-simulant un échantillon de 2 millions de DPE de l'ADEME. Les données de performances de l'enveloppe et des systèmes sont reprises directement des DPE existants, tandis que les métrés sont extraits des géométries de la BDNB. Le calcul est mené sur le moteur de calcul 3CL de la BDNB développé pour le calcul massif. Les résultats sont présentés dans la figure 2. La figure 2b permet de valider le niveau de dispersion causé par l'utilisation des métrés de la BDNB, avec 51% de classes identiques et 92% à ± 1 classe. La figure 2a illustre la préservation de la distribution des consommations en énergie primaire tout en atténuant largement les pics dus aux effets de seuils des classes.



(a) Distributions de consommation



(b) Matrice de confusion

Figure 2 : Validation de l'utilisation du moteur de calcul 3CL-BDNB et des géométries BDNB par comparaison aux DPE ADEME observés.

2.2. MODÈLE GÉNÉRATIF CONDITIONNEL BAT2VEC

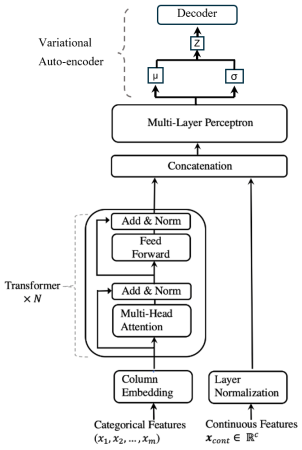


Figure 3 : Architecture du modèle génératif conditionnel bat2vec

Le modèle bat2vec est un réseau de neurones profond dont l'encodeur est inspiré des architectures de type *transformer* tabulaire (Huang et al., 2020; Gorishniy et al., 2021) . Il est combiné avec un auto-encodeur variationnel conditionnel (cVAE) afin d'en faire un modèle génératif stochastique adapté aux tâches de génération de données tabulaires synthétiques (Fonseca and Bacao, 2023) particulièrement adapté à la modélisation de l'incertitude épistémique des modèles de consommation énergétique (Xu et al., 2025). Un schéma d'architecture est présenté figure 3. Les données d'entrées catégorielles (zone géographiques, type de chauffage...) sont encodées par plongement (*categorical embedding*). Les données continues destinées à être reconstruite (coefficient de transmission thermique, ratio de surfaces vitrées...) sont discrétisées puis traitées comme des grandeurs catégorielles par la suite. Le tableau 1 donne la liste des variables reconstruites. Les données d'entrée encodées sont ensuite concaténées et passent à travers un encodeur de type *transformer*, avec 6 couches d'attentions successives avant d'arriver à l'espace latent du cVAE. Le décodeur est constitué d'un perceptron multi-couche dont le nombre de sorties est la somme du nombre de modalité des variables reconstruites. Le jeu de

données d'entraînement est un échantillon de 1,5M de DPE localisés au bâtiment et muni des données descriptives de la BDNB. Cet échantillon est sélectionné avec 3 critères : le diagnostic doit être récent et fiable, avoir une forte fiabilité d'association avec la BDNB et être cohérent avec d'autre source de données connues (année de construction...). La tâche d'entraînement consiste à reconstruire les données d'entrée masquées aléatoirement. La fonction de perte est une perte ELBO classique constituée d'un terme de reconstruction (logloss) et un terme de régularisation de l'espace latent, la divergence de Kullback-Liebler. Ces termes sont calculés indépendamment pour les variables masquées et les variables non masquées, avec un taux de masquage des données de 60% et une pondération de la reconstruction des données masquées de 95%-5%. Les performances du modèle sont évaluées par validation croisée sur les données totalement masquées de 73k DPE. Le modèle converge de manière monotone après 3 époques. La figure 4 présente les matrices de confusion et les distributions marginales prédites par le modèle sur l'échantillon de validation.

2.3. PROPAGATION D'INCERTITUDE MONTE-CARLO ET MOTEUR 3CL-DPE

Les données de performance de l'enveloppe et des systèmes plausibles de chaque bâtiment sont partiellement ou totalement reconstruites avec au moins 100 tirages du modèle bat2vec par bâtiment.

L'inférence des données génère 2 milliards de lignes en 4h sur une infrastructure cloud standard équipée de GPU. La simulation Monte-Carlo de ces données par le moteur 3CL est parallélisée sur 30 cœurs et prend environ 20h de calcul.

Variable	**Unité**	**Description**
Enveloppe		
u_[mur/ph/pb]	$W/(m^2.K)$	Coeff. de transmission thermique des murs, planchers hauts et bas
u_baie_vitree / uw	$W/(m^2.K)$	Coeff. de transmission thermique des baies vitrées (Uw)
type_isolation_[mur/ph/pb]	-	Nature de l'isolation (Intérieure, Extérieure, Répartie)
type_fermeture	-	Type de volets ou fermetures associés aux baies
facteur_solaire_baie / sw	-	Capacité de la baie à transmettre la chaleur solaire
traversant	-	Caractère traversant du logement (impact ventilation)
classe_inertie	-	Niveau d'inertie thermique (Légère à Très lourde)
Systèmes		
type_generateur_chauffage	-	Technologie de production (Chaudière, PAC, etc.)
type_installation_chauffage	-	Configuration du système (Individuel ou Collectif)
type_generateur_ecs	-	Technologie de production d'Eau Chaude Sanitaire
type_ventilation	-	Type de système (Naturelle, VMC SF, VMC DF, etc.)

Tableau 1 : Variables descriptives de l'enveloppe et des systèmes inférées pour la simulation DPE-3CL.

2.4. VALIDATION CROISEE DES PERFORMANCES

La performance de la chaîne de simulation complète est évaluée sur l'échantillon de validation dont tous les DPE sont connus. Les performances sont prédites à l'aveugle (en masquant toutes les données connues des DPE). La figure 5 montre la comparaison des consommations en énergie primaire entre les données brutes des DPE ADEME et les prédictions à l'aveugle. On observe une distribution similaire dans les deux cas tout en gommant les effets de seuils.

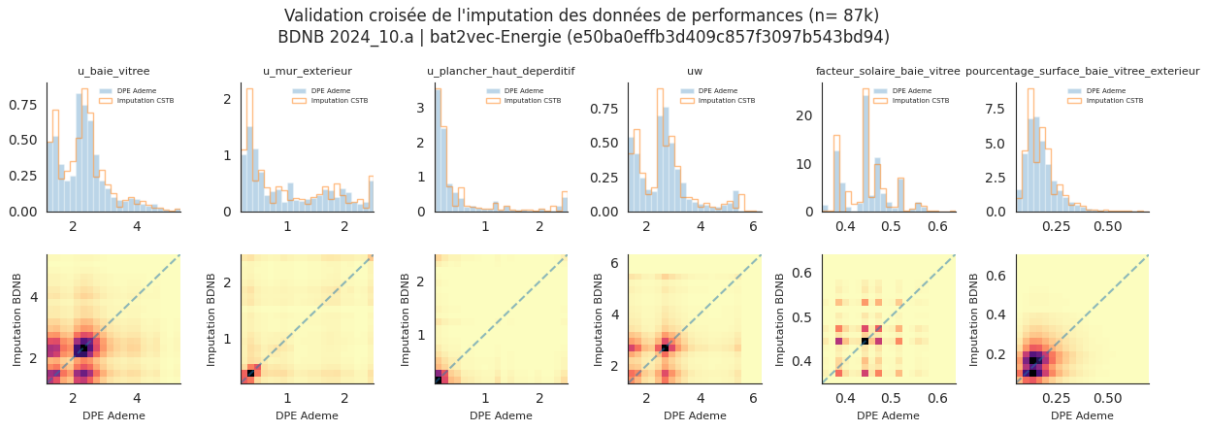


Figure 4 : Validation croisée des données d'entrée des performances de l'enveloppe

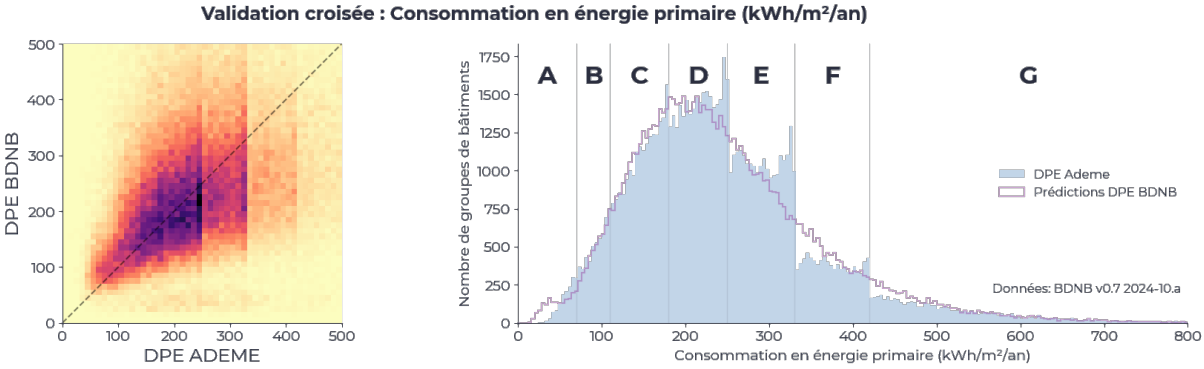
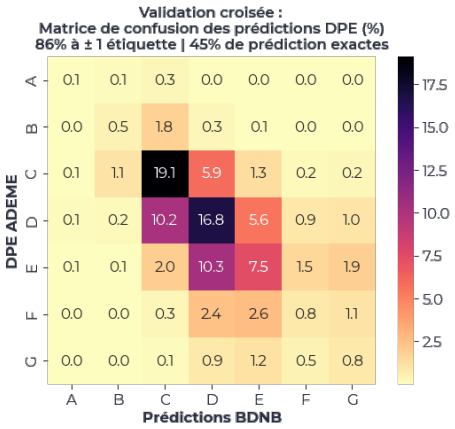
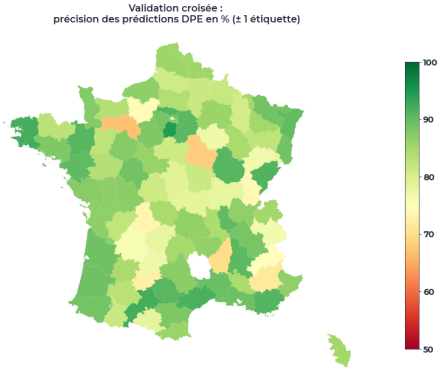


Figure 5 : Validation croisée des prédictions DPE : consommation en énergie primaire.



(a) Validation croisée des classes DPE (arrêté 2021)



(b) Précision des prédictions par département à ± 1 classe DPE

Figure 6 : Estimation de la précision de la chaîne de calcul complète par validation croisée

La figure 6a présente la validation croisée des classes DPE prédites et le niveau de précision à ± 1 classe près par département. La méthode atteint 45% de prédictions de classes exactes et 86% de prédictions correctes à ± 1 classe. Ces performances sont à comparer respectivement

aux 52% et 92% quand les performances DPE sont connues, cf section 2.1.

2.5. ANALYSE DES RÉSULTATS FRANCE ENTIÈRE

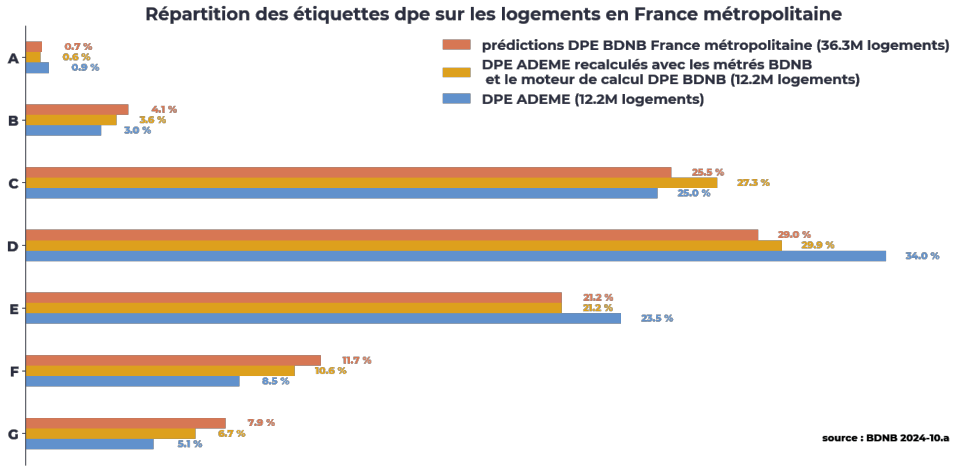


Figure 7 : Distribution des classes DPE : base des DPE ADEME, recalcul des DPE ADEME avec les géométries et le moteur 3CL BDNB, prédictions BDNB France entière.

La figure 7 présente la distribution des classes pour trois échantillons : les DPE bruts de l'ADEME (12,2M), leur recalcul avec les géométries/moteur BDNB, et l'extrapolation à la France entière hors DROM (36,3M de logements). Ce dernier calcul exploite les DPE existants et utilise bat2vec pour imputer les données manquantes (par exemple, les coefficients de planchers pour un DPE d'étage intermédiaire).

Le calcul France entière (hors DROM) couvre 36,3M de logements et exploite les performances connues quand les DPE sont disponibles et utilise le modèle bat2vec pour reconstruire tout ou partie des données manquantes pour le reste du parc. La reconstruction partielle concerne la prédiction des données dans le cas où les données DPE sont incomplètes pour décrire l'ensemble du bâti. Par exemple pour un bâtiment collectif qui serait muni d'un seul DPE sur un étage intermédiaire et dont les coefficients de transmission thermique des planchers haut et des planchers bas sont inconnus.

Les écarts entre les DPE bruts et les recalculs (12,2M de logements) s'expliquent principalement par la correction des effets de seuils qui a tendance à augmenter la proportion de F et de G par rapport aux DPE ADEME (10,6% et 6,7% contre respectivement 8,5% et 5,1% pour les DPE bruts) et diminuer la proportion des étiquettes D et E (29,9% et 21,2% contre respectivement 34% et 23,5% pour les DPE bruts). (Civel et al., 2025)

Les écarts de classes entre ce segment de parc (logements disposant d'un DPE) et l'ensemble des logements s'explique par un biais de sélection de la collecte du DPE connue de la littérature (Chlond et al., 2021). L'analyse des données brutes de l'ADEME montre que la performance du parc évolue par segment. Le parc le plus performant est celui des logements sociaux, suivi par le parc des logements occupés par leurs propriétaires puis le parc locatif privé et enfin le parc des logements vacants.

3. CAS D'ÉTUDE : GISEMENT DE RÉNOVATION URBAINE

3.1. DÉFINITION DES SCÉNARIOS DE RÉNOVATION GLOBALE

Pour évaluer les gisements d'économies d'énergie et de réduction d'émission de gaz à effets de serre, un scénario de rénovation globale dit "maximum" est appliqué virtuellement au parc. Ce scénario impose les exigences minimales du label BBC Rénovation sur l'enveloppe (isolation thermique par l'extérieur avec un coefficient U cible de 0.21 W/m²/K, isolation des planchers hauts et bas). Concernant les systèmes, l'hypothèse centrale est une transition systématique vers des pompes à chaleur (air/eau ou air/air) pour le chauffage et des systèmes thermodynamiques pour l'eau chaude sanitaire.

3.2. ÉVALUATION DU GISEMENT CARBONE

Le gisement carbone est ici défini comme la réduction des émissions annuelles de gaz à effet de serre (GES) en phase d'exploitation, consécutive à une rénovation globale. Son évaluation s'appuie sur le calcul DPE méthode 3CL (arrêté 2021) et couvre les cinq postes réglementaires (chauffage, refroidissement, production d'eau chaude sanitaire, éclairage et auxiliaires). Ce périmètre correspond aux émissions directes et indirectes liées à l'énergie (Scopes 1 et 2) mais ignore l'impact carbone des composants liés aux travaux de rénovation (Scope 3). Ce gisement est obtenu en calculant le différentiel d'émissions entre l'état initial et l'état rénové simulés. De plus, grâce à la re-propagation des échantillons Monte-Carlo sur le scénario final, l'estimation de ces gains est assortie d'un indicateur de fiabilité pour chaque bâtiment.

Au niveau national, si l'ensemble du parc résidentiel bénéficiait de cette rénovation globale, la diminution théorique des émissions annuelles atteindrait 72 Mt de CO_2 eq. Pour référence la SNBC indique que les émissions annuelles du secteur du bâtiment (résidentiel et tertiaire) en 2019 s'élevait à 98 Mt de CO_2 eq (Ministère de la Transition écologique, de la Biodiversité et des Négociations internationales, 2021). L'analyse révèle une forte concentration de ce potentiel : la moitié de ce gisement total (soit 36 Mt de CO_2 eq annuelles) serait accessible en ciblant prioritairement les 20% des logements présentant les gains absolus les plus importants. Il est essentiel de noter que ces logements ne se limitent pas exclusivement aux étiquettes DPE F ou G, la performance énergétique réglementaire étant une métrique relative à la surface, tandis que le gisement absolu dépend fortement de la taille du bâti et du contenu carbone des vecteurs énergétiques.

La figure 8 illustre ce gisement théorique total ventilé par année de construction, en mettant en évidence (en rouge) la fraction accessible via la priorisation de ces 20% de logements présentant le gain absolu le plus important. Spatialement, la figure 9 cartographie ce potentiel à l'échelle du territoire national. Sur cette carte, chaque pixel représente le gisement moyen par mètre carré de surface habitable. On y observe une concentration dans les périphéries urbaines développées durant les Trente Glorieuses, au sein d'un tissu résidentiel majoritairement composé de maisons individuelles occupées par leurs propriétaires.

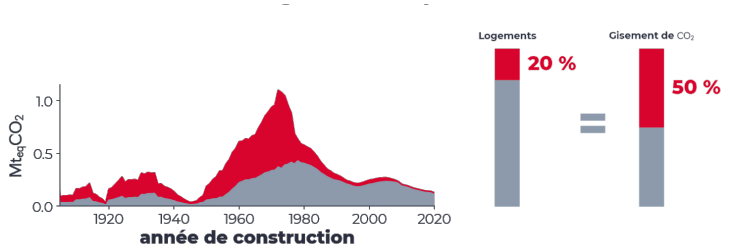


Figure 8 : Gisement de réduction des émissions annuelles de gaz à effet de serre en phase d'exploitation de la rénovation globale du parc par année de construction des bâtiments

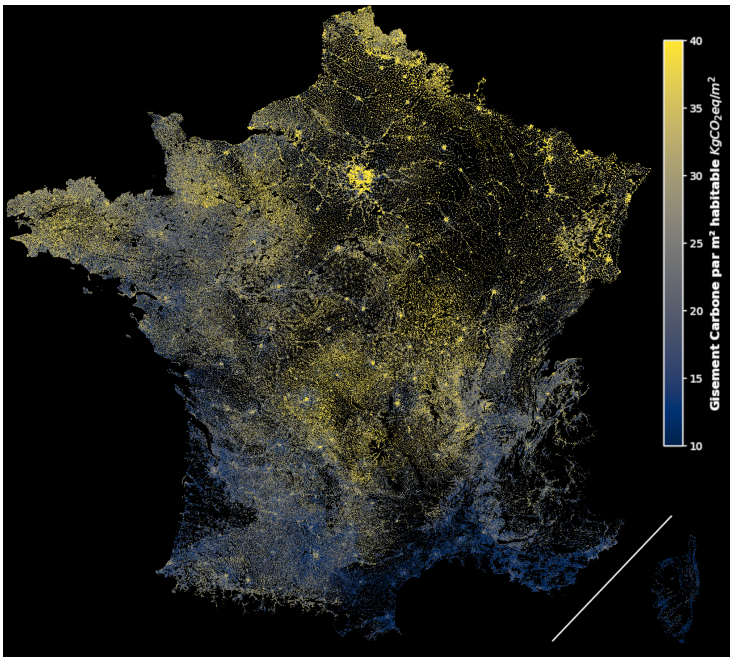


Figure 9 : Gisement de réduction des émissions annuelles de gaz à effet de serre phase d'exploitation de la rénovation globale du parc par m^2 habitable

4. CONCLUSION

La planification de la rénovation énergétique à l'échelle nationale et territoriale nécessite de disposer d'une vision exhaustive de la performance du parc à toutes les échelles. Cet article a présenté une méthodologie probabiliste robuste, couplant un modèle génératif profond (bat2vec) aux géométries de la Base de Données Nationale des Bâtiments (BDNB), afin d'alimenter massivement un moteur de calcul physique réglementaire (3CL).

En substituant une approche déterministe par une propagation d'incertitudes via Monte-Carlo, nous avons pu reconstituer un jumeau énergétique du parc résidentiel français complet (36,3 millions de logements). Les résultats démontrent non seulement la capacité de l'approche à corriger les effets de seuil inhérents aux DPE déclaratifs, mais aussi son aptitude à cibler précisément les politiques publiques : Notre cas d'étude révèle que cibler prioritairement les 20% de logements les plus impactants (construction des Trente Glorieuses, surreprésentation des maisons individuelles occupées par leurs propriétaires) permettrait de capter la moitié du gisement national de réduction d'émissions.

Ces prédictions sont mises à jour à chaque nouveau millésime de la BDNB depuis 2022, intégrant continuellement les nouveaux DPE et l'évolution du parc. Elles sont mises à disposition au grand public via le service Go-Rénove, qui a comptabilisé plus d'un million de visites en 2025, avec près de 8 millions de consultations des prédictions à l'échelle bâtiment.

Cette modélisation physique exhaustive du parc permet l'intégration de nouveaux indicateurs liés à l'adaptation au changement climatique, telles que la prédiction de la vulnérabilité des bâtiments face à la multiplication des vagues de chaleur calculées sur les climats d'aujourd'hui et de demain (Hostein et al., 2026; Toesca, 2026), et l'estimation de l'exposition des personnes vulnérables.

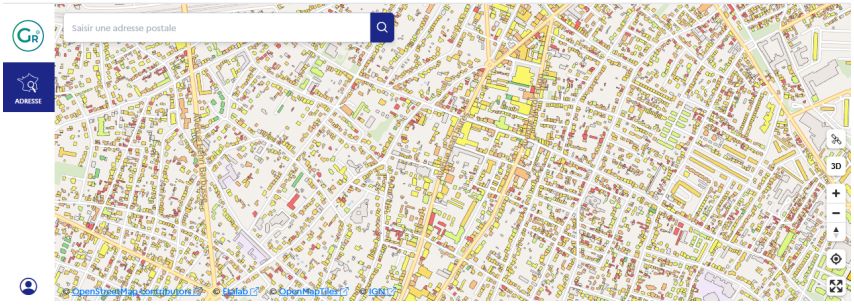


Figure 10 : Classes DPE prédites. Interface de l'application Gorenove

REMERCIEMENTS

Les auteurs remercient les partenaires et financeurs du programme **PROFEEL**, soutenu dans le cadre du dispositif des **Certificats d'Économies d'Énergie (CEE)**. Ces travaux ont été réalisés dans le cadre des projets **GO-RÉNOVE 1 et 2**, portés par le CSTB et l'AQC, visant à fiabiliser et accélérer la rénovation énergétique du parc de bâtiments français.

BIBLIOGRAPHIE

Chlond, B., Gavard, C., and Jeuck, L. (2021). How to support residential energy conservation cost-effectively : An analysis of public financial schemes in france. Discussion paper, ZEW - Leibniz Centre for European Economic Research.

Civel, E., Creti, A., Fack, G., and Herrera-Araujo, D. (2025). Certification, manipulation and competition : evidence from energy performance certificates. Working Paper 2025-04, Chaire Économie du Climat, Paris, France.

Fonseca, J. and Bacao, F. (2023). Tabular and latent space synthetic data generation : a literature review. *Journal of Big Data*, 10 :115.

Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A. (2021). Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34, pages 18932–18943. Curran Associates, Inc.

Hostein, M., Toesca, A., Machard, A., Kyriakodis, G., and Merlet, Y. (2026). Validation and improvement of urban heat island modelling using field measurements. In *Conférence IBPSA France – Lyon – 2026*, Lyon, France.

Huang, X., Khetan, A., Cvitkovic, M., and Karnin, Z. (2020). Tabtransformer : Tabular data modeling using contextual embeddings. arXiv preprint arXiv :2012.06678 (Amazon AWS).

Kyriakodis, G.-E. (2025). Automated generation of urban weather files for building energy simulation. In *Îlots de Chaleur Urbain : Mesures, Impacts et Perspectives Conference*, Nice, France.

Merlet, Y., Toesca, A., Bacon, C., Schetelat, P., and Machard, A. (2025). A simulation-based overheating indicator for 20 million dwellings nationwide. hal-04960739.

Ministère de la Transition Écologique (2025). Arrêté du 13 août 2025 modifiant le coefficient de conversion de l'énergie finale en énergie primaire de l'électricité pour le diagnostic de performance énergétique. `https://www.legifrance.gouv.fr/`.

Ministère de la Transition écologique, de la Biodiversité et des Négociations internationales (2021). Émissions de gaz à effet de serre du secteur du bâtiment en France (scopes 1 et 2). Indicateurs de suivi de la Stratégie Nationale Bas-Carbone (SNBC). Consulté le 4 mars 2026.

Toesca, A., M. Y. H. M. M. A. (2026). Comparaison d'indicateurs de confort thermique estival et de risque sanitaire à la chaleur par simulations sur le parc de bâtiment français. In *Conférence IBPSA France – Lyon – 2026*, Lyon, France.

Xu, X., Hu, Y., Atamturktur, S., Chen, L., and Wang, J. (2025). Systematic review on uncertainty quantification in machine learning-based building energy modeling. *Renewable and Sustainable Energy Reviews*, 218 :115817.