# Modèle de prédiction des DPE à l'adresse V1

Marc Grossouvre, Benoît Génot, Safia Raouf

URBS

Urban Retrofit Business Services

# Modèle de prédiction des DPE à l'adresse V1

January 18, 2023

urbs.fr

Marc Grossouvre, janvier 2023

## Avant-propos

**En rouge, les informations essentielles**

Ceci est une information essentielle.

**En bleu, les informations utiles et accessibles à tous**

Ceci est une information pour tous.

**En vert, les informations qui concernent le DPE d'après juin 2021**

À partir de juin 2021, un nouveau DPE a été instauré.

**Les informations en dehors des cadres de couleur** présentent des informations plus techniques pour les personnes qui découvrent le sujet.

**Texte réglementaire**

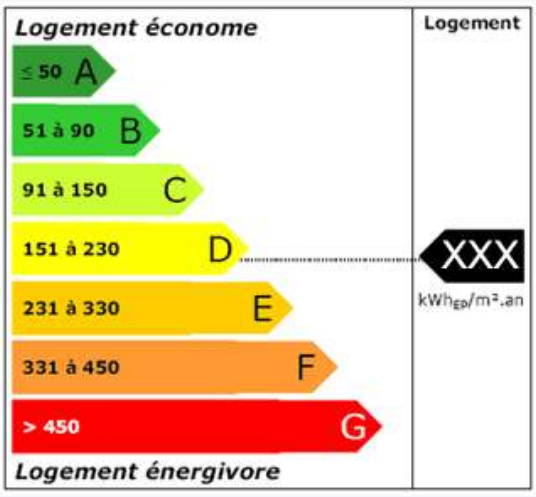
"Le diagnostic de performance énergétique d'un bâtiment ou d'une partie de bâtiment est **un document** qui comprend la quantité d'énergie effectivement consommée ou estimée, exprimée en énergie primaire et finale, pour une utilisation standardisée du bâtiment ou de la partie de bâtiment et **une classification en fonction de valeurs de référence afin que les consommateurs puissent comparer et évaluer sa performance énergétique."**

(Code de la construction et de l'habitat, art. L134-1 à L134-5)

**AVERTISSEMENT**

La présente documentation concerne le DPE tel que réalisé avant juin 2021. Seules quelques indications concernant le DPE selon la réglementation d'après juin 2021 sont disponibles dans les encadrés verts.

**Visuel réglementaire**



# 1 Le diagnostic de performance énergétique (DPE)

## 1.1 Que mesure le DPE ?

**Le DPE, une consommation réelle ou théorique ?**

Le DPE regroupe à la fois des consommations réelles (DPE sur facture pour les bâtiments construits avant 1948) et des valeurs théoriques (DPE sur modèle). Il sert à comparer les performances des logements entre eux.

**Le DPE est obligatoire**

Un DPE doit être fourni par le vendeur d'un logement à l'acheteur et par le bailleur au locataire. Il a une durée de validité de 10 ans.

**Le diagnostiqueur**

Le diagnostic est réalisé par un technicien habilité, le diagnostiqueur. Il est chargé de collecter les informations disponibles sur le logement et de faire les relevés nécessaires. Son travail est contraint par l'information disponible. S'il manque d'information il peut utiliser des **paramètres par défaut**. S'il manque trop d'information, il délivrera un **DPE vierge**.

**Le bâtiment**

La méthode est différente selon que le diagnostic demandé concerne une **partie de bâtiment** (un appartement par exemple), un **bâtiment entier** (une maison par exemple) ou un **immeuble** avec un nombre important de logements. Dans ce dernier cas, on parlera d'**audit énergétique** basé sur un échantillon de logements.

**Le résultat du diagnostic**

Le résultat du diagnostic est exprimé par 2 nombres : une consommation d'énergie par mètre carré et par an et une émission de gaz à effet de serre par mètre carré et par an. Le calcul est réalisé à l'aide d'un logiciel validé par le Ministère de la Transition Ecologique. Ces deux nombres sont associés à des étiquettes entre A et G, dites **étiquette DPE** et **étiquette GES**. Les étiquettes DPE A et B indiquent un bâtiments performant tel qu'un **bâtiment basse consommation (BBC)**. Les étiquettes DPE F et G signalent une **passoire énergétique**.

**Unités** La consommation d'énergie du DPE est exprimée en $kWh/m^2/an$. La quantité de gaz à effet de serre émis est calculée en $kgCO_2/m^2/an$. Ces unités de mesure "favorisent" les grands logements par rapport aux petits et les logements compacts (volume proche du cube) par rapport aux logements qui ont une plus grande surface d'échange avec l'extérieur (maison en "L" par exemple).

**Energie finale, énergie primaire** L'énergie finale d'un logement est la quantité d'énergie délivrée au niveau du compteur (équivalent à ce qui figure sur une facture d'électricité par exemple). L'énergie primaire intègre le rendement de conversion et de transport de l'énergie (centrales, lignes électriques, ...). Ainsi l'électricité a un facteur de conversion de 2.58, c'est à dire que:

$$\text{énergie primaire} = 2.58 \times \text{énergie finale}$$

**La nouvelle étiquette DPE est composite**

Depuis 2021, l'étiquette DPE est déterminée uniquement sur méthode. Elle dépend à la fois de la consommation d'énergie primaire et des émissions de gaz à effet de serre. Le plus pénalisant des deux détermine l'étiquette. Cela pénalise les logements chauffés au gaz.

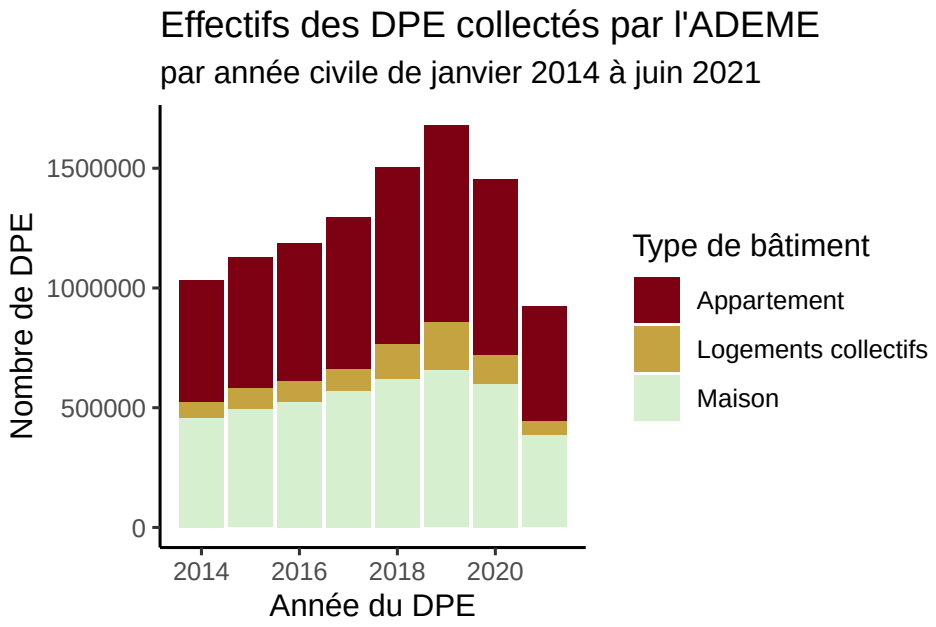
**Modification des facteurs de conversion**

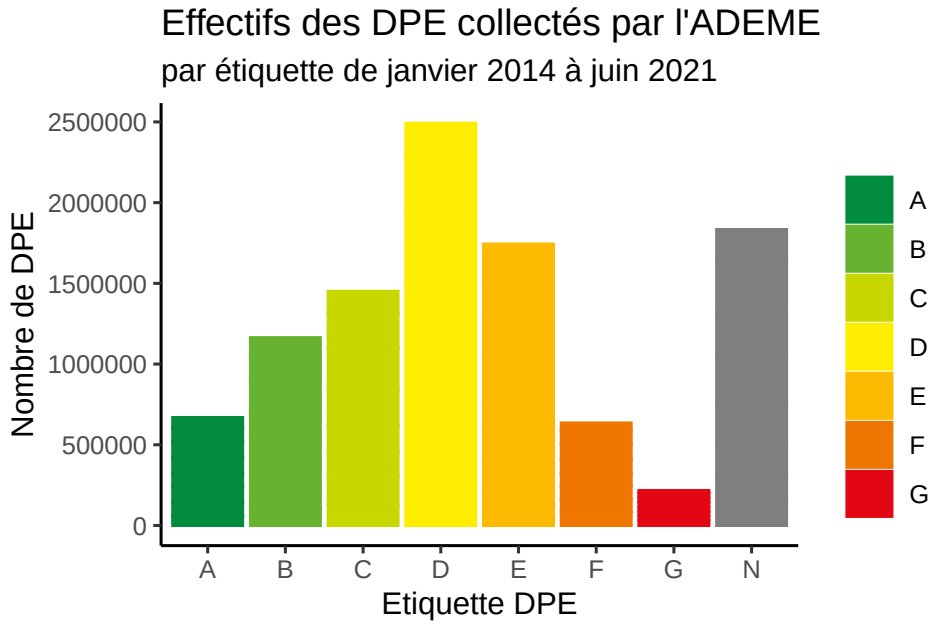
Depuis 2021, le facteur de conversion de l'électricité a été abaissé à 2.30. Du fait de ce changement réglementaire, les logements chauffés à l'électricité ont maintenant une meilleure étiquette DPE.

## 1.2 La collecte des DPE par l'Agence de la transition écologique (ADEME)

**Des données abondantes mais incomplètes**

Depuis 2014, les diagnostiqueurs doivent transmettre à l'ADEME les données des diagnostics énergétiques : mesures, paramètres et résultats (art. L126-32 du Code de la Construction et de l'Habitation). En 2020, les données de ces millions de DPE ont été **ouvertes** après **anonymisation**. Ces données sont difficiles à exploiter car certaines informations comme les adresses sont mal, peu ou pas renseignées.



## Un échantillon non-représentatif

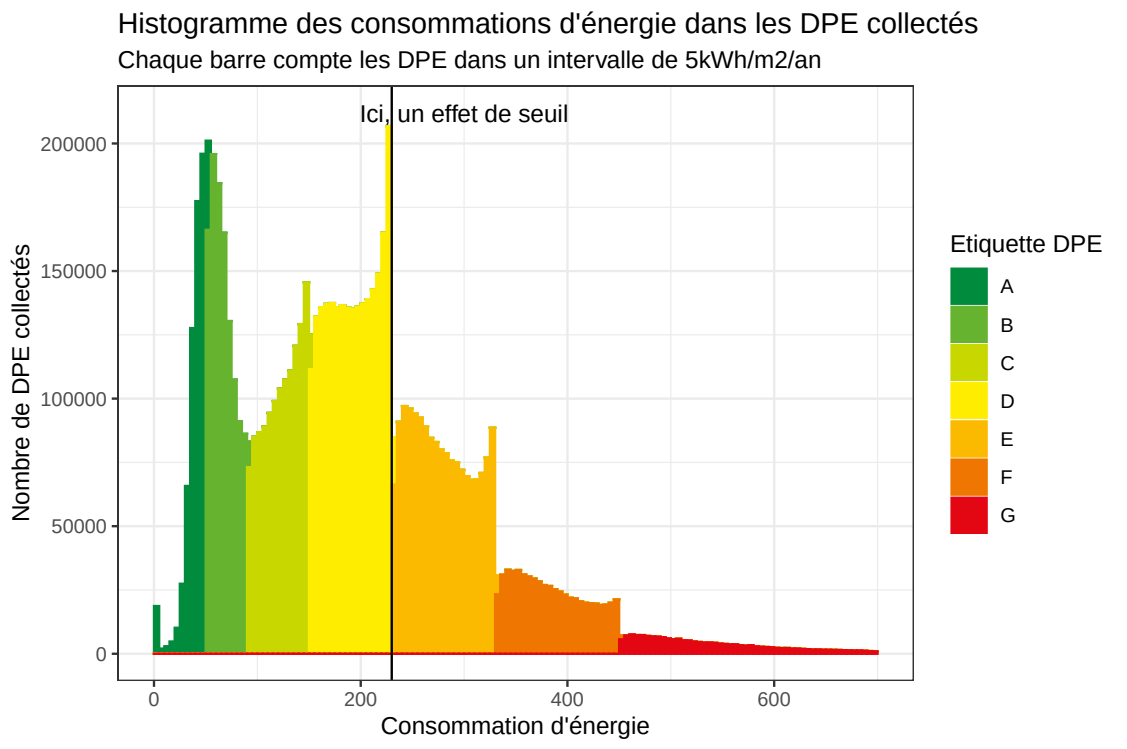
Les données présentent dans la base des DPE ne sont pas représentatives du parc de logement francais car :

- Un logement peut avoir été diagnostiqué plusieurs fois au fil des années ;
- Les logements neufs sont sur-représentés par rapport aux logements anciens ;
- Les immeubles à chauffage collectif, qui ont fait l'objet d'un DPE obligatoire entre 2012 et 2017, sont sur-représentés.

## Combien de passoires thermiques en France ?

Répondre à cette question est difficile, en particulier parce que les DPE collectés ne forment pas un échantillon représentatif. Il faut "redresser" l'échantillon selon certains critères comme l'année de construction du logement ou le type de bâtiment. Plusieurs méthodes de redressement sont possible. On se heurte aux imprécisions et limites de la base ADEME. Par exemple, l'année de construction du logement indiquée dans la base ADEME est souvent arrondie à la décennie voir plus. La question des effets de seuil (voir ci-dessous) pose aussi problème : Doit-on essayer de les gommer ? **Un fait est certain, il y a trop de passoires thermiques en France**.

## Effets de seuil




L'accumulation des DPE juste avant une valeur frontière entre 2 étiquettes s'appelle un effet de seuil.

## 2 Prédire le DPE à l'adresse, pourquoi ?

### Comment massifier la rénovation ?

Les mesures incitatives de type subvention ne suffisent pas. La complexité d'un chantier de rénovation reste en effet un obstacle qui décourage trop de propriétaires. Il paraît donc nécessaire d'aller vers ces propriétaires hésitants, d'être ainsi pro-actif pour acter la décision et les accompagner dans le projet de rénovation.

### Les besoins des collectivités

Pour qu'une approche pro-active soit possible, les collectivités ont besoin de connaître au mieux les logements/bâtiments :

- qui sont des passoires énergétiques ;
- qui sont potentiellement indignes ;
- dont les propriétaires peuvent bénéficier d'aides ;
- dont les occupants sont en situation de précarité énergétique.

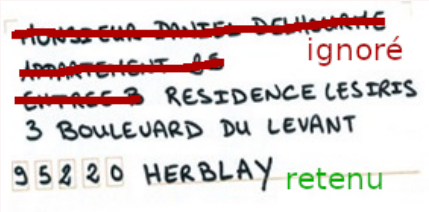
### Les échelles utilisées précédemment

Les approches développées dans la littérature scientifique pour prédire le DPE à grande échelle sont souvent des diagnostics de territoire. Il s'agit d'estimer la répartition des DPE à une échelle statistique : pays et/ou régions pour l'enquête TREMI [3] par exemple, commune [2], quartier [5]. Mais savoir que 20% des logements du quartier sont des passoires énergétiques ne permet pas d'aller rendre visite à leurs propriétaires. Pour cela il faut localiser ces passoires à l'adresse.

### L'adresse comme échelle d'action

U.R.B.S. encourage une approche pro-active en fournissant un DPE observé ou estimé à l'adresse. L'adresse est une donnée qui ne relève pas des données personnelles. L'adresse permet cependant aux collectivités de retrouver dans leurs données contraintes les informations nécessaires pour entrer en contact avec les propriétaires.

### Quel niveau de détail pour l'adresse ?




Seule l'adresse du bâtiment est importante et accessible en open data.

**Le dernier DPE à l'adresse** Comme un même bâtiment (dans notre cas, une même adresse) peut être l'objet de plusieurs DPE collectés par l'ADEME, U.R.B.S. a choisi de prendre comme référence le dernier DPE réalisé à l'adresse. C'est *a priori* le plus à jour et celui qui prend en compte des éventuelles rénovations. Pour une maison, la pertinence de ce choix ne fait aucun doute. Mais pour un immeuble, la situation est parfois plus complexe car on a tantôt un diagnostic pour le bâtiment entier, tantôt un diagnostic pour des logements isolés dont la performance peut-être variable (dernier étage, étage intermédiaire, rez-de-chaussée en particulier).

# 3 Prédire le DPE à l'adresse, comment ?

## 3.1 Les approches existantes

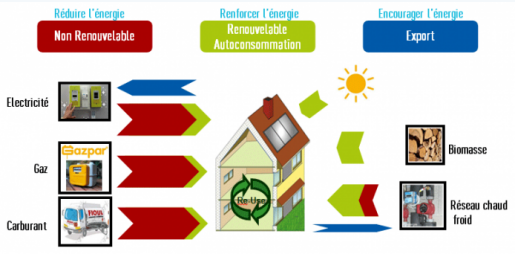
**Modèle physique :**
**Efficace mais très gourmand en données**

Un modèle physique pour prédire l'efficacité énergétique du bâtiment nécessite de connaître un grand nombre de détails sur sa structure. Les matériaux, le type d'isolant et son épaisseur, la conductivité des parois, l'emplacement des radiateurs sont autant d'informations dont les modèles physiques ont besoins. Ces données ne peuvent être connues en détails qu'à condition qu'un diagnostiqueur ait pu étudier le bâtiment. Connaître précisément les caractéristiques physiques de tous les bâtiments en France est impossible aujourd'hui. Pour contourner ce problème, plusieurs équipes de recherche développent des modèles physiques simplifiés. Ces modèles fonctionnent avec moins de paramètres, réduisant ainsi le coût d'acquisition des informations.

**Machine learning :**
**Prometteur et plus sobre en data**

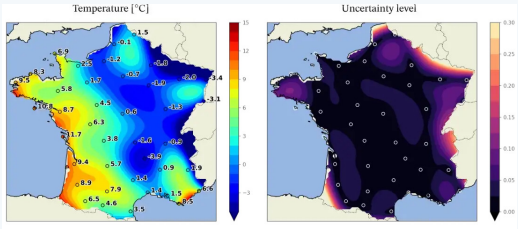
Un algorithme d'apprentissage automatique (machine learning) peut compenser, au moins partiellement le manque d'information sur la structure du bâtiment en utilisant des informations liées (corrélées) à la performance énergétique et disponibles à grande échelle. On sait par exemple que les logements sociaux ont été plutôt bien rénovés et ont une bonne performance énergétique. Cette information n'est pas valorisée dans un modèle physique mais peut-être utilisée avec le machine learning. C'est intéressant car on sait où se trouvent les logements sociaux. De plus, on peut utiliser le fait que le DPE est une information géolocalisée. Bien souvent, des bâtiments proches les uns des autres ont fréquemment des structures similaires mais cette information n'est pas utilisée dans les modèles physiques simplifiés.

**Modèle physique : Bilan thermique d'un bâtiment à énergie positive**




Source ADEME

**Machine learning en géostatistiques : Calcul de la température en France par krigeage**




Source Charles Vanwynsberghe in Medium

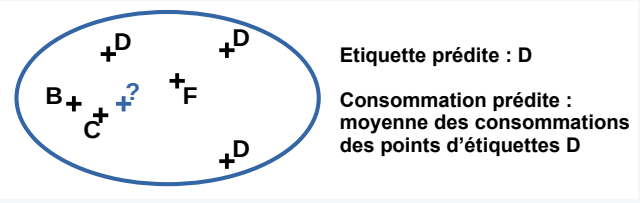
**Quelques références dans la littérature scientifique** La prédiction de l'efficacité énergétique des bâtiments à une échelle statistique utilise souvent des typologies de bâtiment [1]. Les approches plus récentes ont développé des modèles physiques simplifiés [2]. Ces derniers sont parfois hybridés avec des algorithmes de machine learning [4]. Enfin d'autres modèles utilisent des algorithmes de machine learning sans modèle physique. U.R.B.S. s'incrit dans cette dernière démarche pour produire les données présentes dans le logiciel Imope.

## 3.2 Le modèle utilisé dans Imope pour prédire les données DPE

### Le modèle KNN

Le modèle des k plus proches voisins ou k nearest neighbours (KNN) est utile parce qu'il permet d'estimer la performance énergétique d'un bâtiment en utilisant les performances connues des bâtiments **similaires**. La similarité est une proximité géographique, de structure ou de situation socio-économique des occupants/propriétaires. Pour **compenser** l'absence de données techniques sur les bâtiments, nous intégrons au modèle KNN des **variables connexes** susceptibles d'influencer la performance.

### Les plus proches voisins




L'étiquette énergétique prédite pour le bâtiment considéré correspond à l'étiquette majoritaire des k voisins. Nous avons testé k entre 1 et 30 voisins, la valeur optimale est souvent proche de 15.

### Imputer les valeurs manquantes

Une source importante d'information utilisée pour prédire le DPE dans Imope est celle des fichiers fonciers. Mais cette base n'est pas tout à fait complète. Il manque par exemple l'année de construction pour un petit pourcentage de logements. Il est alors nécessaire d'attribuer une valeur pour que le modèle KNN fonctionne. Nous attribuons la médiane des années de construction des bâtiments de la parcelle ou section cadastrale associée. Dans le cas d'une variable booléenne (de type vrai/faux), nous attribuons la valeur majoritaire parmi les bâtiments de la parcelle ou de la section cadastrale associée.

### Un critère d'optimisation

Nous utilisons la balanced accuracy, qu'on peut traduire par "justesse équilibrée". C'est une mesure de qualité de la prédiction qui est focalisée sur les bonnes prédictions.
Point fort : Donne le même poids à toutes les étiquettes.
Point faible : Ne tient pas compte de la gravité des erreurs.

**Choisir une métrique** Pour identifier les plus proches voisins, il faut une métrique adaptée. Nous avons testé 5 transformations de variables : variables brutes, variables ré-échelonnées entre 0 et 1, variables centrées et de variance rendue unitaire, variables de rangs, variables de rangs normalisés. Pour chacune, nous avons testé 3 distances : distance Euclidienne, distance de Manhattan et distance de Hassanat. La distance de Hassanat appliquée aux rangs donne le plus souvent les meilleurs résultats.

### Optimiser pour chaque région

La prédiction des DPE est optimisée séparément pour chaque région de France. Cela permet d'une part de s'assurer de la qualité des prédictions quelle que soit la nature du territoire : les critères de prédiction des DPE ne sont pas les mêmes en Ile de France qu'en Nouvelle Aquitaine.

### Sélectionner les variables

Nous faisons une sélection des meilleures variables pour chaque région de France. Les variables les plus souvent sélectionnées sont les matériaux (murs et/ou toiture), l'énergie de chauffage, l'âge du bâtiment, le type de propriétaire, le type d'occupant. On a donc un mélange entre des caractéristiques physiques et des informations socio-économiques.

### Un échantillon représentatif

Les DPE collectés par l'ADEME ne sont pas représentatifs des bâtiments d'habitation en France. Mais nous devons estimer la performance de l'algorithme sur un échantillon représentatif pour que la performance annoncée soit celle constatée par l'utilisateur. Un échantillon de référence représentatif des bâtiments français est construit en tenant compte de la période de construction (selon la réglementation thermique) et du type maison/logement collectif. C'est sur cet échantillon que la performance est estimée.

**Pré-traitement des données** Après l'intégration des données sources qui alimentent Imope au sein d'une même base cohérente. Une table qui rassemble les données disponibles sur chaque bâtiment est créée. Chaque ligne représente un bâtiment identifié par son adresse. Pour prédire les DPE, on s'intéresse aux colonnes :
$id, idpar, idsec$ Identifiants d'adresse, de parcelle support du bâtiment et de section cadastrale.
$V$ Ensemble des variables numériques issues des fichiers fonciers, des données INSEE et autres.
$typehab$ Variable dans $V$ qui prend la valeur $0$ pour les maisons $1$ pour les appartements.
$regth$ Variable dans $V$ qui prend des valeurs entières de 1 à 9 selon la période de construction du bâtiment, période délimitée par les réglementations thermiques : 1948, 1974, 1982, 1988, 2000, 2005, 2007, 2012.
$dpe$, $ndpe$ Etiquette DPE et, consommation surfacique annuelle associée. **Pour les lignes appariées à la base ADEME, le DPE est présent, sinon, le DPE est manquant.**

**Algorithme 1 :** Prédiction des DPE dans Imope.

**pour** *chaque région de France* $R$ **faire**
- **Données :**
- $X$ table de tous bâtiments de $R$ issue du pré-traitement.
- $(T_u)_{u\in\{1,...,5\}}$, $(D_m)_{m\in\{1,2,3\}}$ Les 5 transformations de variables et les 3 distances testées.
- $M(x, D, k, v)$ Le modèle KNN de données d'apprentissage $x$, utilisant la distance $D$, le nombre de voisins $k$, les variables numériques $v$.
- $BA(M)$ La balanced accuracy d'un modèle $M$ calculée par validation croisée de type leave-one-out.
- `/* 1 Imputation des valeurs manquantes pour les variables de` $V$`. */`
- **pour** $C \in V$ **faire**
  - **si** *la colonne* $X[, C]$ *est incomplète* **alors**
    - **pour** *chaque élément* $X[i, C]$ *manquant* **faire**
      - $X[i, C] := \text{moyenne}(X[j, C] \text{ tel que } X[j, idpar] = X[i, idpar])$
      - **si** $X[i, C]$ *est toujours manquant* **alors**
        - $X[i, C] := \text{moyenne}(X[j, C] \text{ tel que } X[j, idsec] = X[i, idsec])$
- `/* 2 Transformation des variables de` $V$ `*/`
- **pour** $u \in \{1, ..., 5\}$ **faire**
  - $X_u := X$
  - **pour** $C \in V$ **faire** $X_u[, C] := T_u(X_u[, C])$
- `/* 3 Echantillonnage */`
- Soit $X' := F_0(X)$ le filtre qui extrait de $X$ les lignes où $dpe$ est présent.
- Soit $X'' = F_1(X')$ un filtre qui permet d'extraire un échantillon de $15\,0000$ lignes de $X'$ qui soit représentatif de $X$ pour les variables $typehab, regth$.
- Soit $\tilde{X} = F_2(X')$ un filtre indépendant de $F_1$ qui permet d'extraire un échantillon de $50\,0000$ lignes de $X'$ qui soit représentatif de $X$ pour les variables $typehab, regth$.
- **pour** $u \in \{1, ..., 5\}$ **faire**
  - $X'_u = F_0(X_u)$
  - $X''_u = F_1(X'_u)$
- `/* 4 Optimisation */`
- **pour** $u \in \{1, ..., 5\}$ **faire**
  - **pour** $m \in \{1, 2, 3\}$ **faire**
    - **pour** $k \in \{1, ..., 30\}$ **faire**
      - Réaliser une sélection de variables de type "forward selection", de critère $BA$ sur les modèles $M(X''_u, D_m, k, v),\ v \subset V$.
      - L'ensemble des variables finalement sélectionnées est noté $V_{u,m,k}$.
- Le modèle optimal est : $\tilde{M} = M(X''_{\tilde{u}}, D_{\tilde{m}}, \tilde{k}, \tilde{v})$ tel que :
  $BA(\tilde{M}) = \min\{BA(M(X'_u, D_m, k, V_{u,m,k}) \text{ tel que } u \in \{1, ..., 5\},\ m \in \{1, 2, 3\},\ k \in \{1, ..., 30\})\}$.
- `/* 4 Validation du modèle optimal et prédiction des DPE */`
- La balanced accuracy validée du modèle $\tilde{M}$ est $BA(\overline{M})$ où $\overline{M} = M(F_2(X'_{\tilde{u}}), D_{\tilde{m}}, \tilde{k}, V(\tilde{u}, \tilde{m}, \tilde{k}))$.
- L'ensemble des DPE de $X$ est finalement produit en appliquant le modèle $\overline{M}$ à $X_{\tilde{u}}$.

# 4 Pour quels résultats ?

## Résultats à l'échelle du territoire français

**Matrice de confusion pour 10 000 adresses en France**

| Valeurs ADEME | Valeurs prédites | | | | | | |
|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G |
| A | 88 | 40 | 32 | 84 | 42 | 12 | 1 |
| B | 38 | 109 | 114 | 161 | 60 | 15 | 1 |
| C | 25 | 44 | 503 | 863 | 222 | 48 | 1 |
| D | 24 | 35 | 377 | 1930 | 869 | 201 | 7 |
| E | 9 | 20 | 111 | 1084 | 1113 | 310 | 14 |
| F | 2 | 8 | 27 | 318 | 497 | 214 | 12 |
| G | 1 | 3 | 5 | 75 | 144 | 76 | 8 |

Lecture : Sur 10 000 adresses, 377 adresses diagnostiquées D dans la base ADEME sont prédites C.

| Prédictions correctes | Part de toutes les prédictions en % |
|---|---|
| +/- 0 étiquettes | 39.7 |
| +/- 1 étiquettes | 82.9 |
| +/- 2 étiquettes | 95.5 |

**Balanced accuracy**

**29%**

## Variations de performances d'une région à l'autre

| Territoire | Score d'optimisation | Erreur absolue moyenne DPE | RMSE DPE | Erreur absolue moyenne ndpe | RMSE ndpe |
|---|---|---|---|---|---|
| Centre Val de Loire | 35.00 | 0.73 | 1.08 | 70.00 | 97.80 |
| Pays de Loire | 29.86 | 0.83 | 1.19 | 70.47 | 96.24 |
| Ile de France | 29.14 | 0.77 | 1.12 | 74.82 | 103.52 |
| Occitanie | 28.71 | 0.76 | 1.14 | 62.44 | 87.33 |
| Grand-Est | 28.43 | 0.77 | 1.09 | 75.16 | 102.18 |
| Hauts de France | 26.14 | 1.00 | 1.37 | 92.28 | 121.90 |
| Bourgogne Franche-Comté | 26.00 | 0.84 | 1.18 | 80.86 | 108.23 |
| Bretagne | 25.71 | 0.80 | 1.16 | 68.82 | 92.79 |
| Normandie | 25.57 | 0.76 | 1.11 | 72.16 | 98.48 |
| Auvergne Rhône-Alpes | 25.43 | 0.89 | 1.26 | 82.77 | 110.92 |
| Provence Alpes Côte d Azur | 24.29 | 0.80 | 1.15 | 66.26 | 91.14 |
| Nouvelle Aquitaine | 23.43 | 0.98 | 1.36 | 82.81 | 110.69 |
| **FRANCE** | **28.80** | **0.83** | **1.20** | **75.73** | **103.38** |

Lecture : Il y a en moyenne 0,73 étiquettes de différence entre le DPE prédit et le DPE observé dans le Centre-Val de Loire.

Ce tableau appelle plusieurs remarques. D'abord, certaines régions font des contre-performances en terme de RMSE sur le DPE : Hauts de France, Nouvelle Aquitaine, Auvergne-Rhône-Alpes et dans une moindre mesure Pays de Loire. Or ce sont justement les régions qui ont sélectionné les plus petits nombres de voisins (moins de 10). Au contraire, les régions qui ont de plus petites erreurs moyennes ont sélectionné un plus grand nombre de voisins. Pour mémoire, on teste toutes les possibiliés de nombres de voisins entre 1 et 30. Il semble apparaître un optimum autour de 15 voisins.
À noter que les données concernant la Corse sont à ce stade insuffisantes pour implémenter l'algorithme.

# Développements à venir et recherche en cours

**Conclusion**

Nous avons un modèle robuste et pertinent dont la précision doit être améliorée. En effet, si la détection des bâtiments de type BBC est correcte, la détection des passoires énergétiques pour les bâtiments non appariés aux données ADEME reste difficile. Le prochain modèle tiendra compte des résultats de ces études en fixant par avance les variables d'apprentissage, le paramètre $k$, la distance et la transformation des variables. Par contre, nous réaliserons une pondération des variables optimisée pour chaque région. Livraison prévue courant 2023.

**Développements à moyen terme**

En collaboration avec l'Ecole des Mines de Saint-Etienne, nous développons un nouveau modèle géostatistique de prédiction des DPE afin de tenir compte de la granularité des données. Un pre-print est disponible ici.

**Prendre en compte les DPE nouvelle formule**

La livraison 2023 des DPE prédits dans Imope restera apprise des DPE produits avant 2021. Mais par la suite, le stock des nouveaux DPE étant suffisamment important et stable, nous mettrons en place un apprentissage adapté.

La meilleure façon de prédire l'avenir, c'est de le créer.

*Peter Drucker (1909-2005), économiste*

Né en Autriche, mort aux Etats-Unis, il a enseigné le management à la New York University pendant 20 ans. Il est connu pour sa théorie de l'innovation systématique.

## Acronymes et expressions anglaises

**ADEME** Agence de la transition écologique. 4, 5, 6, 7, 8, 11, 12

**BBC** bâtiment basse consommation. 3, 11, 12

**DPE** diagnostic de performance énergétique. 3, 4, 5, 6, 8, 9, 10, 12

**KNN** k plus proches voisins ou k nearest neighbours. 8, 12

**machine learning** apprentissage automatique. 7, 12

## Glossaire

**algorithme de machine learning** Un algorithme de machine learning est une technique de calcul générale qui peut éventuellement être appliquée à l'estimation de la performance énergétique. Un grand nombre d'algorithmes utilisant des approches très variées ont été développés ces dernières décennies. Certains sont issus des mathématiques appliquées, d'autres sont inspirés des phénomènes naturels (algorithmes génétiques, essaims particulaires, réseaux de neurones). On utilise aussi la notion d'intelligence artificielle pour parler de ce domaine, particulièrement lorsqu'il s'agit de travailler sur le langage . 12

**balanced accuracy** Mesure de la qualité d'une prédictions qui consisite, dans le cas du DPE, à faire la moyenne des pourcentages de bonne prédiction par étiquette :

$$\frac{1}{7} \sum_{x \in \{A,B,C,D,E,F,G\}} \frac{\text{nombre de bonnes prédictions } x}{\text{nombre de valeurs observées } x}$$

C'est un nombre entre 0 et 1 que l'on cherche à maximiser . 8, 12

**information géolocalisée** L'information géolocalisée, telle que définie par le décret n° 2011-127 du 31 janvier 2011 relatif au Conseil national de l'information géolocalisée regroupe les "données dont l'information de localisation est essentielle et qui peuvent être géoréférencées" . 7, 12

**modèle physique simplifié** Pour appliquer un modèle physique, il faut collecter typiquement plusieurs dizaines de paramètres associés aux mesures et caractéristiques des matériaux du bâtiment étudié. L'acquisition de ces paramètres est longue, coûteuse et soumise à des aléas de mesure. Développer un modèle physique simplifié, c'est rechercher un modèle qui nécessite moins de paramètres afin de simplifier la procédure de diagnostic énergétique . 7, 12

**modèle physique** Un modèle physique est une technique de calcul de la performance énergétique élaborée par des ingénieurs thermiciens. Il s'agit de faire un bilan énergétique du bâtiment : On calcule les pertes de chaleur et donc la quantité d'énergie qu'il faut apporter pour maintenir une température standard dans les logements. On ajoute à cela la production d'eau chaude, l'éclairage et autres consommations annexes. Eventuellement, la climatisation peut aussi être prise en compte . 7, 12

**RMSE** Le root mean squared error (racine carrée de l'erreur quadratique moyenne) est la distance euclidienne entre les valeurs observées et les valeurs prédites :
$RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(x_i^{obs} - x_i^{pred})^2}$ . 10, 12

## References

[1] Ilaria Ballarini, Vincenzo Corrado, Francesco Madonna, Simona Paduos, and Franco Ravasio. Energy refurbishment of the Italian residential building stock: energy and cost analysis through the application of the building typology. *Energy Policy*, 105:148–160, June 2017.

[2] Alessio Mastrucci, Paula Pérez-López, Enrico Benetto, Ulrich Leopold, and Isabelle Blanc. Global sensitivity analysis as a support for the generation of simplified building stock energy models. *Energy and Buildings*, 149:368–383, August 2017.

[3] Jean-Philippe Rathle. Les réductions des émissions de gaz à effet de serre liées aux rénovations. Résultats de l'enquête TREMI 2020. Technical report, Observatoire national de la rénovation énergétique, September 2022.

[4] Pascal Schetelat, Lucie Lefort, and Nicolas Delgado. Urban data imputation using multi-output multi-class classification. *Building to Buildings: Urban and Community Energy Modelling*, November 2020.

[5] Wenwen Zhang, Caleb Robinson, Subhrajit Guhathakurta, Venu M. Garikapati, Bistra Dilkina, Marilyn A. Brown, and Ram M. Pendyala. Estimating residential energy consumption in metropolitan areas: A microsimulation approach. *Energy*, 155:162–173, July 2018.