Exploration des Comportements Énergétiques Résidentiels : Identification des Facteurs Clés et Perspectives pour l’Optimisation

Kévin Tang

Mohamed-Essaid Khanouche

Parisa Ghodous

Août 2024

Université Claude Bernard Lyon 1



**Résumé**. La consommation énergétique des ménages représente une part significative de la consommation énergétique globale. Pour approfondir notre compréhension de la production, de la gestion et du stockage des excédents d’énergie, il est crucial d’élucider les schémas implicites des comportements des consommateurs et d’identifier les facteurs influençant leurs performances. Cet article vise principalement à analyser de manière descriptive le modèle de consommation énergétique des ménages en utilisant les données du RECS2020 [1], en mettant l’accent sur l’impact du comportement sur la consommation d'énergie. À cette fin, nous nous concentrons sur la sélection du sous-ensemble de caractéristiques le plus pertinent à partir d’un large ensemble de données, ce qui permet une meilleure compréhension des informations, une réduction du temps de calcul et une amélioration des performances prédictives. Les résultats de cette étude peuvent aider les décideurs à examiner les comportements des familles à divers niveaux de la société pour s’assurer que leur qualité de vie est adéquate ou nécessite des améliorations.

**Mots-clés:** Feature selection - Smart home - Household energy consumption - Correlation analysis

# **Sommaire**

[**Sommaire 3**](#_sr7c106oktkx)

[**1. Introduction 4**](#_hxsr4p7b4f5r)

[**2. Travaux connexes 5**](#_8obrolxg0qqi)

[**3. Définition du problème 6**](#_pqrx0jo7quju)

[**3. Méthodologie proposée 6**](#_urcet1nvncnz)

[3.1 Présentation du dataset 7](#_hn0b8h4g5rs2)

[3.2 Prétraitements 7](#_vhvdqzbv6h9y)

[3.2.1 Suppression des variables concernant les énergies autres que l’électricité 7](#_h9c292i5q4o4)

[3.2.2 Suppression des variables indicatrices d’imputation et de calibration 8](#_ip3ji9y7ul7k)

[3.2.4 Gestion des valeurs manquantes 8](#_arr6os2u099l)

[3.2.5 Réduction des variables cibles 9](#_mpqsrqk59ldw)

[3.2.6 Réduction du nombre de variables 9](#_ptga0o6sl9ur)

[3.2.7 Réduction du nombre d’échantillons 9](#_7oknh6azls2g)

[3.2.8 Conclusion du pré-traitement 10](#_bkkom789404n)

[3.3 Sélection des variables 10](#_ndiid1z50cuk)

[3.3.1 Corrélation sur l’ensemble des données 10](#_oa6nvlf7npi)

[3.3.2 Corrélation sur les données catégorisées 11](#_hqjdismvl78)

[3.3.3 Gain d’information sur l’ensemble des données 13](#_8tjo4s2pzwyb)

[3.3.4 Gain d’information sur les données catégorisées 14](#_tdi2gu8y5mkj)

[3.3.5 Intersection des 4 méthodes 15](#_4mw8sel8knjb)

[3.4 Analyse descriptive des variables sélectionnées 15](#_ywgwnirghde5)

[3.4.1 Détermination du nombre optimal de clusters 15](#_qzloabjmokzq)

[3.4.2 K-Means 17](#_hb82l8u6z04e)

[3.4.3 Kernel K-Means 18](#_mvvk7bxqdua)

[3.4.4 DBSCAN 19](#_u1mineb8g1m2)

[3.4.3 Analyse des clusters 21](#_c5nxoqv80xgf)

[3.5 Analyse descriptive des variables comportementales sélectionnées 22](#_ttwzph31klbt)

[3.5.1 Détermination du nombre optimal de clusters 23](#_2zq1vu93icns)

[3.5.2 Deux clusters 23](#_2oug6ctsl8yg)

[3.5.3 Trois clusters 24](#_v4ucijge0yb2)

[3.5.3 Six clusters 24](#_1ad7t07cg0xn)

[3.5.4 Dix clusters 25](#_l7zixlt1ikvk)

[3.5.5 DBSCAN 26](#_o08ul9yxrtt6)

[**4. Discussion 27**](#_r355w4su6hmv)

[**5. Conclusion 28**](#_4bdilksimxm9)

[**Références 30**](#_pa4t1a18xbx4)

# **1. Introduction**

La population mondiale a atteint 8 milliards d'habitants en novembre 2022, soit un milliard de plus qu'en 2010, et devrait franchir la barre des 10 milliards avant la fin du siècle selon les prévisions de l'ONU [2,3]. Cette croissance rapide, associée à l'urbanisation et à l'augmentation des besoins en énergie, contribue à une hausse significative de la consommation énergétique des ménages, qui représentent environ un tiers de la consommation énergétique mondiale. En France métropolitaine, les ménages représentent actuellement 36 % de la consommation d’électricité, faisant de ce secteur le plus énergivore en proportion, devant le secteur tertiaire et le secteur industriel [4]. Cette tendance soulève des préoccupations importantes quant à la durabilité énergétique.

L'essor des dispositifs intelligents (IoT), dont le nombre devrait atteindre 55 milliards d'unités d'ici 2025, ouvre de nouvelles perspectives pour le suivi et l'optimisation de la consommation énergétique. Ces dispositifs collectent des quantités massives de données sur les habitudes, les préférences et les comportements des utilisateurs, offrant ainsi une opportunité sans précédent de comprendre et de gérer de manière plus efficace la consommation d'énergie. Le paradigme de l'Internet du Comportement (IoB) répond à ce besoin en permettant une analyse approfondie des comportements des utilisateurs et de leur impact sur la consommation énergétique.

Cependant, le défi reste de taille : comment interpréter ces données pour développer des modèles de gestion d’énergie efficaces dans les espaces résidentiels, en se basant sur le comportement des occupants ? Alors que les approches existantes se concentrent principalement sur la sélection des caractéristiques liées aux appareils utilisés, aux caractéristiques des occupants (comme l'âge, le revenu) ou aux habitations elles-mêmes (superficie, nombre de chambres), notre approche se distingue par un focus sur les caractéristiques comportementales. Nous cherchons à comprendre comment les comportements humains influencent la consommation d'énergie, offrant ainsi une analyse plus fine et une opportunité d'améliorer l'efficacité énergétique globale.

Dans cet article, nous analysons l'ensemble de données RECS2020, qui contient des informations sur la consommation énergétique résidentielle des ménages américains en 2020. Nous développons des modèles d'apprentissage pour identifier les corrélations entre les comportements des utilisateurs et leurs dépenses énergétiques. Notre méthodologie inclut la collecte et le prétraitement des données comportementales, la sélection des caractéristiques pertinentes, et l'application d'algorithmes de clustering pour segmenter les ménages en groupes significatifs. Ces clusters sont ensuite analysés pour extraire des informations sur les habitudes de consommation et pour construire des archétypes comportementaux. Ces archétypes permettent de prédire les profils de consommation les plus énergivores, offrant ainsi des insights précieux pour le développement de stratégies énergétiques adaptées et ciblées.

# **2. Travaux connexes**

Dans cette section, nous passons en revue les travaux connexes dans trois domaines principaux : la prévision de la consommation d’énergie, les méthodes d’extraction de caractéristiques et les techniques permettant d’influencer les comportements de consommation d’énergie.

La thèse de Mouna Labiadh \cite{ref\_article4} présente une nouvelle méthodologie de prédiction de la consommation énergétique des bâtiments, en particulier lorsque les données historiques font défaut. Cette approche utilise un modèle MLP siamois combiné à KNN et à divers modèles temporels comme SVR, MLP, LSTM, CNN et Seq2seq, en se concentrant principalement sur les caractéristiques physiques des bâtiments. D'autres chercheurs ont également exploré des modèles prédictifs avec des approches différentes. Wang et al. \cite{ref\_article7} ont développé un réseau neuronal convolutionnel profond (DCNN) basé sur ResNet, intégrant des caractéristiques temporelles et météorologiques pour des prévisions précises. Chan et Yeo \cite{ref\_article8} ont utilisé un transformateur clairsemé, qui offre une précision comparable aux RNN tout en étant plus rapide. Hadjout et al. \cite{ref\_article9} \cite{ref\_article10} ont utilisé une combinaison de modèles LSTM, GRU, TCN et SARIMA, améliorés par des techniques d'apprentissage d'ensemble. Syed et al. \cite{ref\_article11} ont exploré diverses architectures LSTM, telles que les couches bidirectionnelles et unidirectionnelles. Alhussein et al. \cite{ref\_article12} ont intégré des architectures hybrides qui combinent CNN et LSTM pour une meilleure précision des prévisions. Enfin, Li et al. \cite{ref\_article13} ont développé le modèle Trans-T2V, combinant un transformateur avec Time2Vec pour des prévisions de consommation d'électricité affinées.

Outre les modèles prédictifs, plusieurs études se sont attachées à identifier les facteurs clés influençant la consommation énergétique. Par exemple, une étude \cite{ref\_article1} a analysé la consommation énergétique des ménages américains à l'aide de l'ensemble de données RECS 2015, bien qu'elle se soit principalement concentrée sur les facteurs socio-économiques plutôt que sur les aspects comportementaux. Une autre étude \cite{ref\_article3} a identifié des facteurs liés au mode de vie tels que l'utilisation de la climatisation, les pratiques de lessive et l'utilisation de l'ordinateur personnel, en se basant sur les données RECS de 2001 et 2005. Comme notre étude, le travail de \cite{ref\_article2} visait à construire des archétypes comportementaux pour mieux comprendre les habitudes de consommation énergétique dans le secteur résidentiel. Les données, tirées du projet ENERGIHAB de l'Agence nationale de la recherche, comprenaient 35 variables couvrant l'hygiène, l'alimentation, le chauffage, l'éclairage, les pratiques de loisirs et l'occupation du logement auprès de 1 363 ménages d'Île-de-France. Cette étude a réussi à identifier sept archétypes comportementaux distincts. Cependant, la portée de cette étude est quelque peu limitée par le faible nombre de variables, la focalisation géographique sur l'Île-de-France et l'utilisation de données recueillies par téléphone en 2010, ce qui peut introduire des biais ou un manque d'actualité des informations. Ces études, bien qu'intéressantes, limitent souvent leur analyse aux caractéristiques physiques et sociodémographiques sans approfondir le comportement des occupants.

Les avancées récentes ont exploré des moyens d'influencer et de modifier les comportements de consommation d'énergie. Haya Elayan et al. \cite{ref\_article5} ont proposé un cadre IoB (Internet of Behavior) décentralisé visant à prédire la consommation d'électricité et à influencer le comportement des appareils IoT. Ce cadre, utilisant des données de consommation d'électricité des ménages français et des modèles LSTM, semble prometteur pour réduire la consommation d'énergie en affectant directement le fonctionnement des appareils connectés. Bien que ce cadre ait été testé sur un nombre limité d'appareils, il offre une perspective optimiste sur la possibilité de réduire la consommation d'électricité non seulement au niveau du ménage, mais aussi à celui d'une population entière. Une autre étude d'Elayan et al. \cite{ref\_article6} a combiné l'IoB avec l'IA explicable (XAI) pour optimiser les décisions d'utilisation d'énergie tout en rendant ces décisions transparentes pour les utilisateurs. L'intégration d'un moniteur et d'un contrôleur d'énergie (EMC) et d'un composant « Explainer » aide les utilisateurs à comprendre et à accepter les recommandations d'économie d'énergie, encourageant ainsi un comportement plus durable. Une décision bien expliquée, qui montre clairement les avantages en termes d’économies d’énergie et de réduction des coûts, a le potentiel de motiver les utilisateurs à adopter des pratiques plus durables.

En résumé, la littérature existante propose une variété de techniques avancées pour la prédiction de la consommation énergétique, l’extraction de caractéristiques et l’influence des comportements des appareils connectés. Les modèles prédictifs, tels que ceux basés sur le LSTM ou d’autres réseaux neuronaux, sont déjà bien développés pour anticiper les besoins énergétiques. De même, des méthodes sophistiquées d’extraction de caractéristiques permettent d’analyser les facteurs sociodémographiques et les caractéristiques des bâtiments influençant la consommation. Les techniques d’influence montrent également un potentiel prometteur pour modifier les comportements des utilisateurs et optimiser la consommation d’énergie. Cependant, notre étude se démarque en se concentrant spécifiquement sur l’identification et la segmentation des comportements des ménages, un aspect souvent sous-exploré dans les recherches précédentes. En combinant ces techniques existantes avec une analyse comportementale approfondie, nous proposons d’améliorer l’efficacité des interventions énergétiques. En identifiant avec précision les habitudes quotidiennes qui impactent l’empreinte énergétique, notre approche offre un potentiel accru de réduction de la consommation énergétique de manière ciblée et personnalisée, en adaptant les solutions non seulement aux comportements des occupants eux-mêmes, mais aussi à leurs habitudes de vie. Cette originalité permet de mieux comprendre comment les comportements peuvent être modifiés pour réduire l’empreinte énergétique, contribuant ainsi à des politiques énergétiques plus efficaces et ciblées.

# **3. Définition du problème**

La gestion de la consommation d'énergie domestique est un défi majeur dans le contexte actuel de changement climatique et de transition vers des sources d'énergie renouvelables. Une meilleure compréhension des comportements de consommation d'énergie des ménages peut aider à identifier les leviers d'action pour réduire la demande en énergie, améliorer l'efficacité énergétique et diminuer les émissions de gaz à effet de serre.

Le problème se situe au croisement de plusieurs enjeux : environnemental, économique et social. La consommation d'énergie domestique contribue significativement aux émissions de gaz à effet de serre, ce qui a un impact direct sur le changement climatique. Réduire la consommation d'énergie est essentiel pour atteindre les objectifs de réduction des émissions fixés par divers accords internationaux. Sur le plan économique, l'énergie représente une part importante des dépenses des ménages. En identifiant les comportements énergivores et en proposant des solutions pour les réduire, il est possible de diminuer les factures énergétiques des foyers, ce qui peut avoir un impact positif sur le pouvoir d'achat. Sur le plan social, il existe des disparités importantes dans la consommation d'énergie en fonction des revenus, de la taille du logement et des équipements possédés par les ménages. Comprendre ces disparités est crucial pour proposer des mesures équitables et adaptées à chaque type de ménage.

L'objectif principal de cette étude est de segmenter les ménages en fonction de leurs comportements de consommation d'énergie afin d'identifier les profils de consommation et les facteurs clés influençant cette consommation. Les objectifs spécifiques incluent l'identification et la caractérisation des différents groupes de ménages en fonction de leur consommation d'énergie et de leurs caractéristiques socio-économiques, l'analyse des principaux facteurs de consommation d'énergie au sein de chaque groupe pour comprendre quelles activités ou quels équipements contribuent le plus à la consommation d'énergie, et la proposition de recommandations pour réduire la consommation d'énergie dans chaque groupe identifié, en tenant compte des spécificités de chaque groupe pour maximiser l'efficacité des mesures.

# **3. Méthodologie proposée**

L’objectif de notre recherche est d'identifier les comportements ayant le plus d’impact sur la consommation énergétique dans les ménages, à partir de l’ensemble de données RECS de 2020. Nous avons donc sélectionné trois variables cibles : la quantité totale d’électricité utilisée en kilowattheures (kWh), la quantité totale d’électricité utilisée en BTU (BTUEL), et le coût total de l’électricité consommée en dollars (DOLLAREL).

Dans cette section, nous décrirons d'abord l’ensemble de données utilisé. Ensuite, nous détaillerons les différents prétraitements appliqués à ce dataset, ainsi que les méthodes de sélection des caractéristiques employées pour identifier les comportements les plus significatifs pour notre recherche. Enfin, nous analyserons les résultats obtenus après l’application des algorithmes de clustering et l'étude de corrélations.

## **3.1 Présentation du dataset**

Le RECS (Residential Energy Consumption Survey) est une enquête détaillée menée aux États-Unis, visant à recueillir des informations sur la consommation d'énergie résidentielle des ménages américains. Réalisée par l'Energy Information Administration (EIA), cette enquête est l'une des sources de données les plus complètes et fiables sur l'utilisation de l'énergie dans le secteur résidentiel.

L'édition 2020 du RECS collecte des données sur divers aspects de la consommation énergétique des foyers, y compris les types et quantités de combustibles utilisés, les caractéristiques des logements, les appareils électroménagers, les systèmes de chauffage et de climatisation, ainsi que des comportements récents tels que l’utilisation de l’énergie solaire ou le lieu de recharge des véhicules électriques. L'enquête couvre ainsi un large éventail de 799 variables, permettant une analyse approfondie et multidimensionnelle des habitudes de consommation d'énergie.

Le RECS utilise un échantillon représentatif de 18 496 ménages américains pour s'assurer que les résultats peuvent être généralisés à l'ensemble de la population. Les enquêtes détaillées sont envoyées aux ménages sélectionnés et peuvent être administrées sous forme de questionnaires papier, en ligne, par interviews téléphoniques ou lors de visites sur site.

Pour plus d’informations, vous pouvez consulter le site de l’EIA [1].

## **3.2 Prétraitements**

Pour assurer la qualité et la pertinence des analyses, plusieurs prétraitements ont été appliqués aux données du RECS 2020. Ces étapes sont cruciales pour nettoyer et préparer les données avant l’application des techniques de sélection de caractéristiques et des algorithmes de clustering.

### 3.2.1 Suppression des variables concernant les énergies autres que l’électricité

Étant donné que notre recherche vise à identifier les comportements influençant la consommation d’électricité, nous avons supprimé 117 variables relatives à d'autres types d'énergie, telles que le gaz naturel, le propane, le fioul et le bois. En excluant ces variables non pertinentes, nous avons pu concentrer notre analyse spécifiquement sur les facteurs influençant la consommation d'électricité. Cette démarche permet de réduire la complexité du dataset et facilite l'interprétation des résultats.

### 3.2.2 Suppression des variables indicatrices d’imputation et de calibration

Les variables indicatrices d’imputation signalent les valeurs manquantes ou estimées pour d'autres variables, tandis que les variables de calibration ajustent les poids des réponses pour refléter une population représentative. Afin d’affiner notre analyse et de garantir la pertinence des données, nous avons supprimé 407 de ces variables utilisées par l'EIA. Bien que ces variables soient cruciales pour les processus internes de l’EIA, elles ne sont pas nécessaires pour notre étude, qui se concentre spécifiquement sur les comportements de consommation électrique.

3.2.3 Encodage des variables catégorielles

Pour préparer notre dataset à une analyse approfondie, nous avons converti les variables catégorielles en une forme que les algorithmes de machine learning peuvent comprendre. Nous avons utilisé la méthode LabelEncoder de la bibliothèque scikit-learn. Cette technique transforme chaque valeur catégorielle en une valeur numérique unique, indiquant la catégorie à laquelle elle appartient.

Dans notre dataset, nous avons identifié 7 variables catégorielles nécessitant un encodage. L'application du LabelEncoder à ces variables a permis de convertir les données catégorielles en une représentation numérique tout en préservant l'intégrité et la signification des informations originales.

### 3.2.4 Gestion des valeurs manquantes

Pour assurer la qualité et la complétude de notre dataset, nous avons adopté une méthode pragmatique pour traiter les valeurs manquantes (NaN).

La première étape a consisté à identifier toutes les lignes du dataset contenant une ou plusieurs valeurs manquantes. Nous avons identifié 290 échantillons contenant un NaN, ce qui représente environ 1,57 % du dataset. Étant donné que cette proportion est très faible, nous avons décidé de supprimer ces échantillons afin de simplifier le traitement des données.

Après la suppression des lignes contenant des valeurs manquantes, le nombre de ménages dans notre dataset a diminué, passant de 18 496 à 18 206. Cette approche permet de garantir que notre analyse se base sur un dataset complet, minimisant ainsi les biais et les erreurs potentiels liés à l'imputation des valeurs manquantes.

### 3.2.5 Réduction des variables cibles

Au début de notre étude, nous avions identifié trois variables cibles principales pour évaluer la consommation d'électricité : la quantité totale d’électricité utilisée en kilowattheures (kWh), la quantité totale d’électricité utilisée en BTU (BTUEL), et le coût total de l’électricité consommée en dollars (DOLLAREL). Cependant, après avoir réalisé une analyse de corrélation entre ces variables, nous avons constaté que certaines d'entre elles étaient fortement corrélées. En particulier, la variable BTUEL était fortement corrélée avec kWh, car ces deux mesures sont interconvertibles (1 kWh équivaut à environ 3 412 BTU).

En conséquence, nous avons décidé de réduire le nombre de variables cibles à deux : kWh et DOLLAREL. Cette réduction permet de simplifier l'analyse tout en conservant les mesures essentielles pour évaluer la consommation d’électricité.

### 3.2.6 Réduction du nombre de variables

Dans le cadre du prétraitement des données, nous avons identifié les paires de variables fortement corrélées dans notre dataset pour réduire la redondance. Cette analyse a conduit à la suppression de 69 variables.

Par exemple, nous avons supprimé la variable "state\_postal" et conservé "state\_name" pour éviter la redondance entre les codes postaux des états et leurs noms. De même, nous avons supprimé la variable "TELLWORK", qui indique qu’un membre du ménage télétravaille, au profit de variables plus précises comme "TELLDAYS", "TLDESKTOP" ou "TLLAPTOP", qui apportent plus d’informations sur les modalités du télétravail. Cette approche a également été appliquée à d'autres domaines, tels que l'aide à l'énergie, les véhicules électriques, les lumières et le type de logement, où nous avons choisi de garder les variables fournissant les informations les plus détaillées et pertinentes.

Cette démarche a permis de simplifier le dataset tout en conservant les informations les plus pertinentes pour notre analyse.

### 3.2.7 Réduction du nombre d’échantillons

Les données isolées, souvent considérées comme des points aberrants, peuvent biaiser les modèles d'apprentissage automatique et fausser les résultats des analyses statistiques. Il est donc essentiel de les identifier et de les supprimer pour obtenir des analyses plus précises et fiables, permettant ainsi une meilleure compréhension des données.

Dans notre étude, nous avons utilisé l'algorithme Isolation Forest pour identifier ces données isolées. Isolation Forest isole les anomalies en construisant des arbres d'isolation. Chaque arbre partitionne aléatoirement les données jusqu'à ce que chaque point soit isolé dans une feuille. Les anomalies, étant rares et différentes, sont isolées plus rapidement, ce qui signifie qu'elles nécessitent moins de divisions pour être séparées des autres points. Chaque point de données reçoit un score d'anomalie basé sur la profondeur moyenne à laquelle il est isolé dans les arbres. Plus la profondeur est faible, plus le point est susceptible d'être une anomalie.

Cette méthode a permis de détecter 894 échantillons considérés comme aberrants, représentant environ 5% du dataset. La suppression de ces points a réduit le nombre total de ménages dans notre dataset, passant de 18 206 à 17 312. En éliminant ces données isolées, nous avons significativement amélioré la qualité du dataset, ce qui permet d'obtenir des analyses plus précises et robustes.

### 3.2.8 Conclusion du pré-traitement

À l'issue des étapes rigoureuses de prétraitement, nous avons considérablement affiné notre dataset pour le rendre plus pertinent et exploitable pour notre analyse.

Le nombre d’échantillons a été réduit de 18 496 à 17 312, suite à la suppression de 894 points aberrants identifiés comme isolés et de 290 échantillons contenant au moins une valeur manquante. Cette réduction a permis de maintenir une base de données plus cohérente et représentative.

En ce qui concerne les variables, nous avons simplifié notre dataset en réduisant le nombre total de variables de 799 à 208. Cette diminution a été réalisée en éliminant les variables relatives à d'autres types d'énergie, les variables indicatrices d’imputation et de calibration, ainsi que celles fortement corrélées. Cette démarche a permis de réduire les redondances et de concentrer notre analyse sur les facteurs les plus pertinents.

Ces étapes de prétraitement ont non seulement optimisé la qualité des données, mais ont également facilité des analyses plus précises et significatives. En fin de compte, ces ajustements permettent d’obtenir des résultats plus fiables et d’approfondir notre compréhension des comportements de consommation électrique.

## **3.3 Sélection des variables**

La sélection des variables est une étape cruciale dans le prétraitement des données, visant à améliorer la qualité des modèles prédictifs en identifiant les variables les plus pertinentes. Dans notre étude, nous avons appliqué plusieurs méthodes pour affiner notre choix de variables.

### 3.3.1 Corrélation sur l’ensemble des données

La première méthode de sélection des variables repose sur l’analyse de la corrélation entre toutes les variables du dataset et les variables cibles.

Nous avons d'abord calculé la corrélation entre chaque variable et la variable cible KWH pour identifier celles ayant une relation significative avec la consommation électrique. Les variables dont la corrélation avec KWH était supérieure à la moyenne des corrélations calculées ont été sélectionnées.

De manière similaire, nous avons évalué la corrélation entre chaque variable et la deuxième variable cible, DOLLAREL, pour identifier les variables influençant le coût de la consommation d’électricité. Les variables dont la corrélation avec DOLLAREL était également supérieure à la moyenne des corrélations calculées ont été retenues.

Enfin, les variables pertinentes pour chacune des deux cibles ont été croisées pour obtenir leur intersection, ce qui a permis de sélectionner 76 variables importantes pour les deux variables cibles, garantissant leur pertinence dans les deux contextes d’analyse.

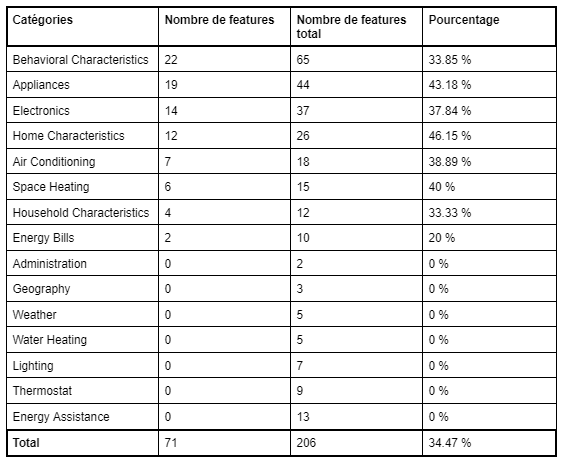
### 3.3.2 Corrélation sur les données catégorisées

La deuxième méthode de sélection des variables repose sur la catégorisation des variables du dataset et l’analyse de la corrélation au sein de ces catégories. Cette approche permet de structurer le processus de sélection en traitant les variables selon leurs catégories spécifiques.

L'EIA a partitionné les variables en 14 catégories distinctes. Nous avons ajouté une catégorie supplémentaire spécifiquement dédiée aux variables liées au comportement, comprenant 88 variables identifiées manuellement. Les 15 catégories sont les suivantes :

* Administration : Cette catégorie regroupe des variables administratives et contextuelles qui fournissent des informations essentielles pour l'identification, la classification climatique et le type urbain des répondants.
* Géographie : Cette catégorie regroupe des variables géographiques qui fournissent des informations essentielles sur la localisation et les caractéristiques géographiques des répondants telles que l'État de résidence.
* Météo : Cette catégorie regroupe des variables météorologiques qui fournissent des informations essentielles sur les conditions climatiques et les caractéristiques météorologiques des zones habitées par les répondants telles que la température extérieure, du sol ou des ampoules. Elle concerne 5 variables.
* Caractéristiques du logement : Cette catégorie regroupe des variables qui fournissent des informations détaillées sur les caractéristiques physiques et structurelles des logements des répondants telles que la taille et le type de logement.
* Caractéristiques du ménage : Cette catégorie regroupe des variables qui fournissent des informations détaillées sur les caractéristiques sociodémographiques et économiques des ménages des répondants telles que le nombre d’habitants, leur niveau d’éducation ou les revenus du ménage.
* Electroménager : Cette catégorie regroupe des variables qui fournissent des informations détaillées sur la possession, l'utilisation et les caractéristiques des appareils électroménagers présents dans les logements des répondants.
* Électronique : Cette catégorie regroupe des variables qui fournissent des informations détaillées sur la possession, l'utilisation et les caractéristiques des appareils électroniques présents dans les logements des répondants.
* Éclairage : Cette catégorie regroupe des variables qui fournissent des informations détaillées sur les types, l'utilisation et les caractéristiques des systèmes d'éclairage présents dans les logements des répondants.
* Climatisation : Cette catégorie regroupe des variables qui fournissent des informations détaillées sur les systèmes de climatisation, incluant leur présence, leur utilisation et leurs caractéristiques dans les logements des répondants.
* Chauffage de l’eau : Cette catégorie regroupe des variables qui fournissent des informations détaillées sur les systèmes de chauffage de l'eau, y compris leur type, leur utilisation et leurs caractéristiques dans les logements des répondants.
* Chauffage de l’espace : Cette catégorie regroupe des variables qui fournissent des informations détaillées sur les systèmes de chauffage, incluant leur type, leur utilisation et leurs caractéristiques dans les logements des répondants.
* Thermostat : Cette catégorie regroupe des variables qui fournissent des informations sur les dispositifs de contrôle de température, y compris leur type, leur utilisation et leurs caractéristiques dans les logements des répondants. Elle concerne 9 variables.
* Factures d’énergie : Cette catégorie regroupe des variables qui fournissent des informations détaillées sur la responsabilité du paiement des factures d’énergie, la présence d’un compteur intelligent ou le lieu de recharge du véhicule électrique.
* Aide à l’énergie : Cette catégorie regroupe des variables qui fournissent des informations sur les aides et subventions reçues pour la consommation d'énergie. Caractéristiques comportementales : Cette catégorie regroupe des variables de toutes les autres catégories qui fournissent des informations sur les habitudes et les comportements des ménages en matière de consommation d'énergie, y compris les routines quotidiennes ou les préférences d'utilisation.

Après le prétraitement, nous avons calculé la corrélation entre les variables et les deux variables cibles, puis sélectionné celles dont la corrélation était supérieure à la moyenne au sein de chaque catégorie. Cette approche a permis de conserver un total de 71 variables.

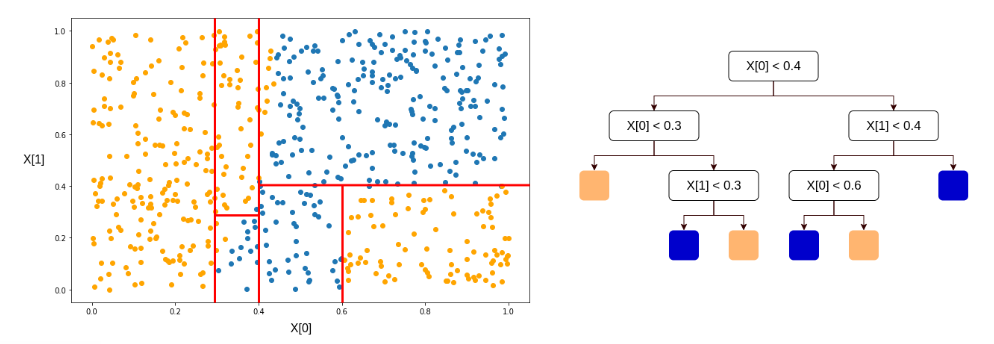


*Figure 1 : Tableau résultat de la méthode 2*

### 3.3.3 Gain d’information sur l’ensemble des données

Le **DecisionTreeRegressor** est un algorithme de régression qui construit un arbre de décision pour prédire des valeurs continues, telles que la consommation d’énergie (KWH) ou le coût (DOLLAREL).

**Fonctionnement de l'algorithme** : L’arbre de décision divise les données en posant des questions successives sur les caractéristiques des données d’entraînement. Pour chaque nouvelle observation, l’arbre suit les branches correspondant aux caractéristiques de l’observation et prédit une valeur basée sur la moyenne des valeurs des données qui arrivent à la feuille terminale.



*Figure 2 : Illustration du fonctionnement de DecisionTreeRegressor*

**Évaluation de l'importance des caractéristiques** : L'algorithme évalue l’importance de chaque caractéristique en mesurant la réduction de l’impureté, quantifiée par l’erreur quadratique apportée par chaque division de l’arbre. À chaque nœud, il sélectionne la caractéristique qui minimise cette impureté, optimisant ainsi les prédictions.

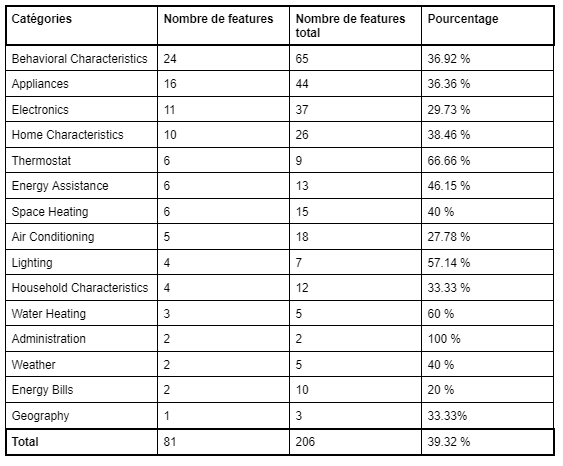
**Sélection des caractéristiques** : Après avoir déterminé l'importance de chaque caractéristique, nous classons ces caractéristiques en fonction de leur contribution à la réduction de l’erreur quadratique globale. Une courbe cumulative est ensuite tracée pour visualiser la contribution de chaque caractéristique. En identifiant le point où la courbe se stabilise, nous pouvons déterminer le nombre optimal de caractéristiques à retenir, équilibrant ainsi précision et simplicité du modèle.

Cette approche permet de sélectionner les variables les plus pertinentes et de construire un modèle de régression efficace, capable de faire des prédictions précises tout en évitant le surapprentissage. Appliquée à l'ensemble des données prétraitées, cette méthode a permis de sélectionner 63 variables.

### 3.3.4 Gain d’information sur les données catégorisées

Comme mentionné précédemment, les données d'origine sont organisées en différentes catégories. Pour chaque catégorie, nous avons appliqué la méthode du coude afin d'identifier le nombre optimal de variables à conserver. Cette approche consiste à tracer la réduction cumulative de l'impureté en fonction du nombre de variables retenues et à déterminer le seuil au-delà duquel l'ajout de variables supplémentaires n'améliore que marginalement la réduction d'impureté.

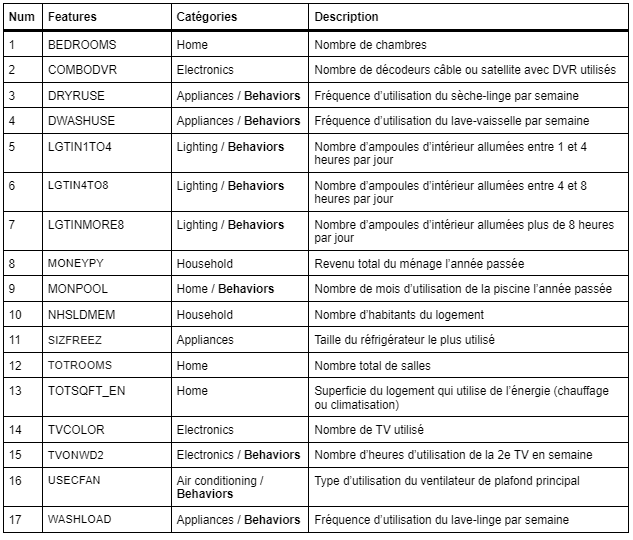
En appliquant cette méthode de manière individuelle à chaque catégorie, nous avons pu sélectionner 83 variables pertinentes. Cette approche permet d'assurer que les variables choisies contribuent de manière significative à la prédiction tout en maintenant la simplicité et l'efficacité du modèle.



*Figure 3 : Tableau résultat de la méthode 4*

### 3.3.5 Intersection des 4 méthodes

L'intersection des résultats obtenus par les quatre méthodes de sélection a permis d'identifier 17 variables parmi les 208 initiales ayant le plus grand impact sur la consommation d'énergie. Cette sélection conjointe met en lumière les variables les plus significatives en combinant les perspectives des différentes méthodes. Parmi ces 17 variables, 9 proviennent des 65 variables comportementales, soulignant leur importance particulière dans l'analyse de la consommation d'énergie.



## **3.4 Analyse descriptive des variables sélectionnées**

À l'issue du prétraitement, nous avons sélectionné 17 caractéristiques présentant la plus forte influence sur les attributs cibles. Étant donné que les données ne sont pas étiquetées, nous avons recours à des méthodes de clustering non supervisées pour analyser ces données.

### 3.4.1 Détermination du nombre optimal de clusters

Pour déterminer le nombre optimal de clusters dans notre analyse, nous avons utilisé quatre méthodes complémentaires : la méthode du coude avec l'inertie, le silhouette score, le score de Davies-Bouldin, et le critère de Calinski-Harabasz. Chaque méthode fournit une perspective différente sur la qualité du clustering et aide à identifier le nombre de clusters le plus approprié pour représenter au mieux la structure des données.

**Méthode du Coude avec l'Inertie** : La méthode du coude est une approche classique pour déterminer le nombre optimal de clusters. Elle repose sur l'inertie, également appelée somme des carrés intra-cluster. L'inertie mesure la dispersion des points de données au sein de chaque cluster : plus l'inertie est faible, plus les points sont proches du centre du cluster. En traçant l'inertie en fonction du nombre de clusters, nous obtenons une courbe qui montre généralement une réduction rapide au début, suivie d'une stabilisation. Le "coude" de cette courbe, où la réduction de l'inertie commence à diminuer de manière significative, indique le nombre optimal de clusters. Ce point représente un équilibre entre la compacité des clusters et la complexité du modèle.

**Silhouette score** : Le silhouette score évalue la qualité du clustering en mesurant la similarité des points au sein de leur propre cluster par rapport à la similarité avec les points des autres clusters. Le score varie de -1 à 1 : un score proche de 1 indique que les points sont bien regroupés dans leur propre cluster et éloignés des autres clusters, tandis qu'un score proche de -1 suggère que les points pourraient être mal classifiés. Pour différents nombres de clusters, nous avons calculé le silhouette score moyen. Le nombre de clusters qui maximise ce score est considéré comme optimal, car il indique la meilleure séparation et la plus grande cohésion entre les clusters.

**Score de Davies-Bouldin** : Le score de Davies-Bouldin est une autre mesure de la qualité du clustering, qui évalue la compacité et la séparation des clusters. Il est calculé en mesurant la moyenne des ratios de similarité entre chaque cluster et le cluster le plus similaire. Un score plus bas indique une meilleure séparation entre les clusters, suggérant que les clusters sont bien distincts et compacts. En comparant les scores pour différents nombres de clusters, le nombre de clusters qui minimise le score de Davies-Bouldin est choisi comme optimal, car il reflète la meilleure séparation des groupes.

**Critère de Calinski-Harabasz** : Le critère de Calinski-Harabasz, également connu sous le nom de "variance ratio criterion", évalue la qualité du clustering en comparant la variance intra-cluster à la variance inter-cluster. Plus ce ratio est élevé, meilleure est la séparation entre les clusters et plus les clusters sont compacts. Nous avons calculé ce critère pour différents nombres de clusters et sélectionné le nombre qui maximise le critère de Calinski-Harabasz. Cela indique que les clusters sont bien définis avec une séparation claire et une cohésion interne élevée.

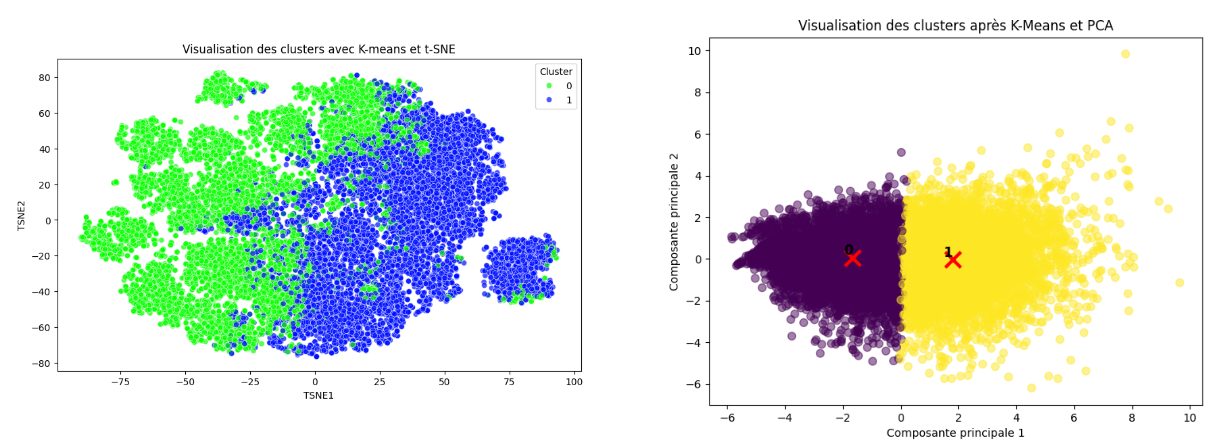
L'application des quatre méthodes de sélection du nombre optimal de clusters a conduit à une conclusion convergente : toutes les méthodes suggèrent que le nombre optimal de clusters est 2. L'accord entre ces différentes méthodes renforce la conclusion selon laquelle deux clusters sont le choix le plus approprié pour capturer la structure sous-jacente des données. Ce consensus multi-méthodes valide la robustesse et la fiabilité de cette solution, garantissant ainsi une représentation adéquate des groupes au sein de notre analyse.

### 3.4.2 K-Means

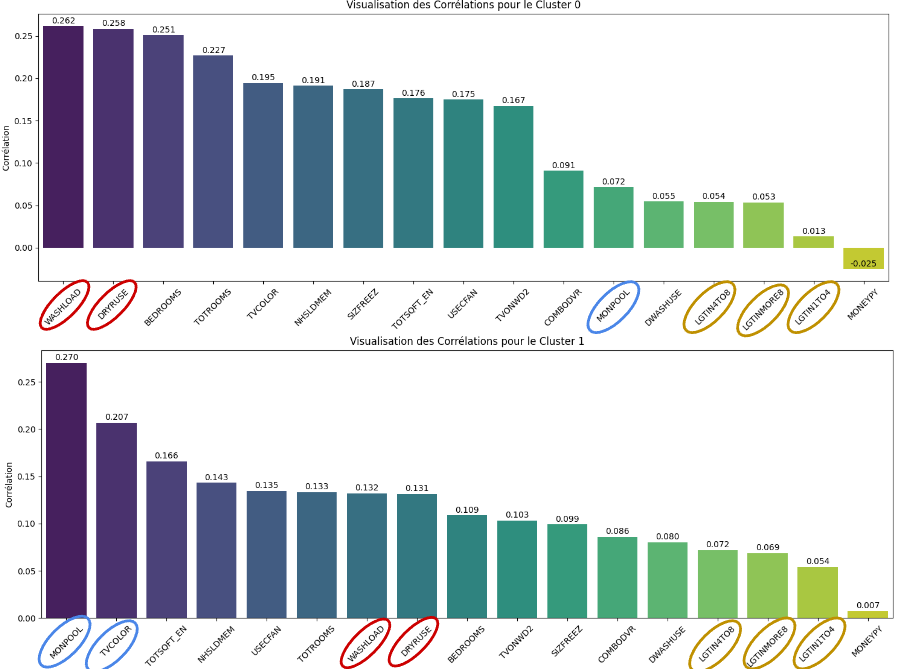
Le K-means est une méthode de clustering non supervisée qui vise à diviser un ensemble de données en un nombre fixe de groupes, appelés clusters. L'objectif est de minimiser la variance au sein des clusters et de maximiser la variance entre les clusters.

L'algorithme K-means fonctionne en plusieurs étapes. L'algorithme commence par l'initialisation des centres des clusters. Ces centres peuvent être déterminés aléatoirement ou en utilisant des techniques spécifiques comme K-means++ pour améliorer la sélection initiale des centres. Chaque point de données est assigné au cluster dont le centre est le plus proche. Cette proximité est généralement mesurée par la distance euclidienne entre les points de données et les centres des clusters. Une fois que tous les points ont été assignés à des clusters, les centres des clusters sont recalculés en prenant la moyenne des points appartenant à chaque cluster. Les étapes d'assignation des points et de mise à jour des centres sont répétées jusqu'à ce que les centres des clusters convergent, c’est-à-dire que les changements deviennent négligeables, ou jusqu’à ce qu’un critère d’arrêt soit atteint (tel qu’un nombre maximal d’itérations). L’algorithme s’arrête donc lorsque les centres des clusters ne varient plus de manière significative, ce qui indique que les clusters ont atteint une forme de stabilité.

Dans le cadre de cette recherche, l'algorithme K-means a été appliqué en utilisant le nombre optimal de clusters déterminé par les méthodes précédentes, établi à deux clusters.



*Figure 4 :* Visualisation du clustering des variables sélectionnées avec K-Means



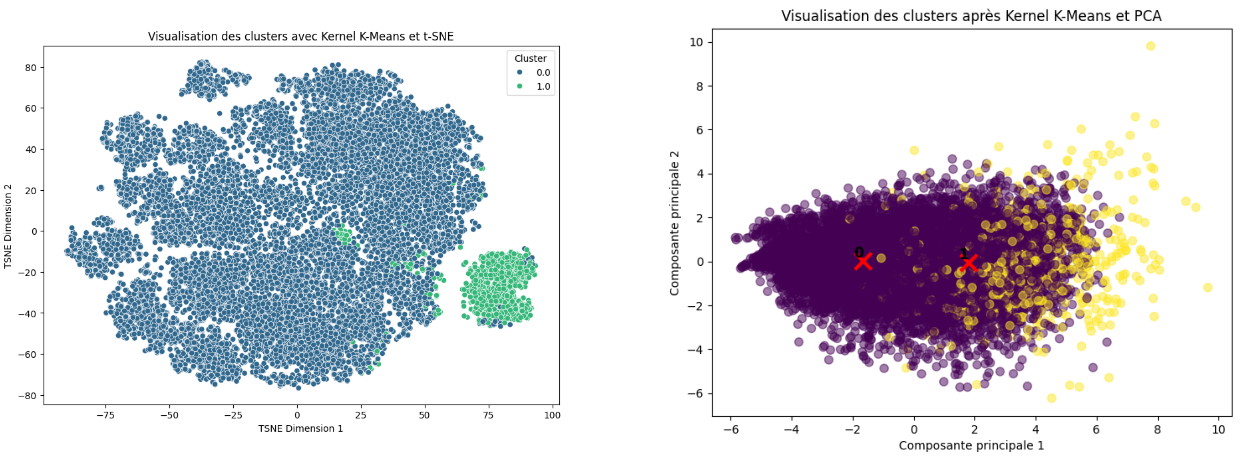
*Figure 5 : Corrélations des Clusters avec K-Means (2 Clusters)*

### 3.4.3 Kernel K-Means

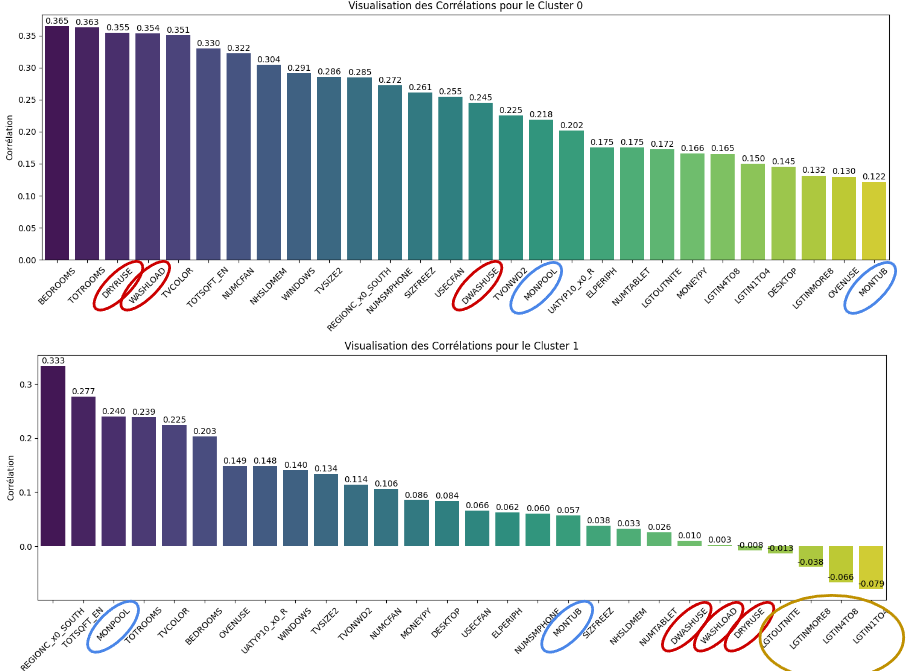
Kernel K-means est une extension de l'algorithme K-means classique qui utilise une fonction noyau pour mapper les données dans un espace de dimension plus élevée, permettant ainsi de capturer des structures non linéaires dans les données. En transformant les données à l'aide d'une fonction noyau, Kernel K-means peut identifier des clusters qui ne seraient pas détectables par K-means classique, car ils se trouvent dans un espace où les frontières entre les clusters ne sont pas linéaires.

L'algorithme suit des étapes similaires à K-means : initialisation des centres de clusters, assignation des points aux clusters en fonction des distances calculées avec la matrice noyau, mise à jour des centres de clusters, et itération jusqu'à convergence. Cependant, au lieu d'utiliser la distance euclidienne traditionnelle, Kernel K-means utilise les distances calculées dans l'espace transformé par la fonction noyau. Cette approche permet de mieux capturer des relations complexes entre les points de données.

Dans notre étude, après avoir appliqué K-means, nous avons utilisé Kernel K-means pour vérifier si des structures plus complexes existaient dans les données. Les résultats ont confirmé que les clusters identifiés étaient robustes et nous ont permis de mieux comprendre les comportements de consommation énergétique des ménages.



*Figure 6 :* Visualisation du clustering des variables sélectionnées avec Kernel K-Means

**

*Figure 7 : Corrélations des Clusters avec Kernel K-Means (2 Clusters)*

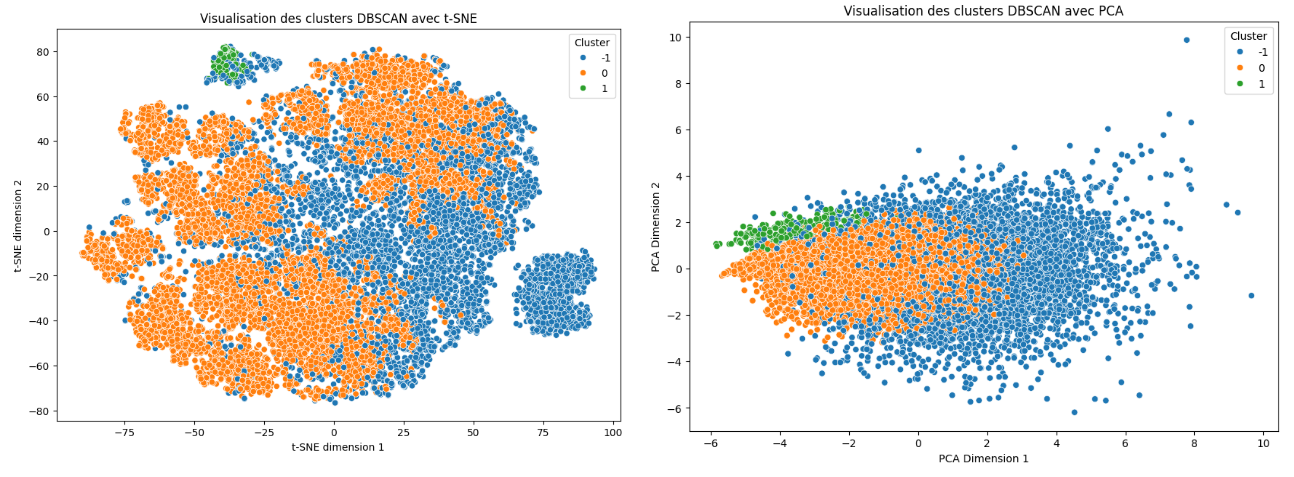
### 3.4.4 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) est un algorithme de clustering non supervisé largement utilisé pour identifier des clusters dans des ensembles de données en se basant sur la densité des points de données. Contrairement à d'autres méthodes de clustering telles que K-Means, qui nécessitent de définir le nombre de clusters à l'avance, DBSCAN peut identifier automatiquement le nombre de clusters en fonction de la densité locale des points. L'algorithme repose sur deux principaux paramètres : ε (epsilon), qui représente la distance maximale entre deux points pour qu'ils soient considérés comme voisins, et minPts, le nombre minimum de points dans le voisinage ε pour qu'un point soit considéré comme un point central (core point).

DBSCAN commence par sélectionner un point non visité dans l'ensemble de données et identifie tous les points dans son voisinage ε. Si le nombre de points voisins est supérieur ou égal à minPts, le point est considéré comme un point central et un nouveau cluster est créé. Si le nombre de points voisins est inférieur à minPts, le point est marqué comme du bruit, bien qu'il puisse être inclus plus tard dans un cluster s'il est un voisin d'un point central. Pour chaque point central, l'algorithme explore tous les points dans son voisinage et, si ces points sont également des points centraux, leur voisinage est également exploré. Ce processus continue jusqu'à ce que le cluster ne puisse plus être étendu. L'algorithme répète ces étapes jusqu'à ce que tous les points soient visités.

DBSCAN présente plusieurs avantages. Il peut identifier des clusters de formes arbitraires, contrairement à K-Means qui suppose des clusters sphériques, et traite explicitement les points de bruit, ce qui améliore la qualité des clusters formés. De plus, il détermine automatiquement le nombre de clusters en fonction de la densité des points, ce qui est avantageux lorsque le nombre de clusters n'est pas connu à l'avance. Il est également efficace pour les bases de données contenant de grands volumes de données et fonctionne bien avec les bases de données de densité variable.

L'application de l'algorithme DBSCAN à notre ensemble de données, avec des paramètres ε définis à 2 et minPts fixé à 19, a révélé la formation de quatre clusters distincts. Cependant, il est important de noter que ces clusters sont relativement petits en taille. La majorité des points de données ont été classés comme des anomalies, reflétant la nature stricte de l'algorithme dans la définition des points de cluster en fonction de la densité locale. Cette caractéristique de DBSCAN souligne sa capacité à identifier et isoler les points aberrants dans les ensembles de données, mais elle peut également indiquer que les paramètres ε et minPts pourraient nécessiter un ajustement pour mieux capturer les structures de clusters potentielles présentes dans les données.



*Figure 8 : Visualisation du clustering des variables sélectionnées avec DBSCAN*

### 3.4.3 Analyse des clusters

Les deux méthodes de clustering, K-means et Kernel K-means, ont segmenté les ménages en deux groupes distincts, chacun mettant en évidence des caractéristiques spécifiques de consommation énergétique.

Les ménages appartenant au premier cluster se caractérisent par des logements plus petits, un revenu plus faible, et une possession et utilisation moindre des équipements. Les principaux facteurs de consommation d'énergie dans ce cluster sont liés aux besoins essentiels d'un logement, principalement l'utilisation d'appareils électroménagers pour des tâches quotidiennes comme la cuisson, la réfrigération et le lavage.

Les ménages du second cluster, en revanche, se distinguent par des logements plus grands, un revenu élevé, et une utilisation beaucoup plus intensive des appareils électroménagers et électroniques. Dans ce cluster, les principaux facteurs de consommation d'énergie incluent des équipements liés aux loisirs, tels que la piscine, qui sont largement responsables de la consommation élevée d'électricité.

**K-Means** a identifié deux clusters distincts. Le premier cluster comprend des ménages avec un revenu moyen compris entre 30 000 et 35 000 dollars par an, une superficie de logement moyen de 125 mètres carrés, et des habitudes de consommation principalement axées sur les besoins essentiels. Le second cluster, en revanche, inclut une large gamme de ménages à revenu élevé, avec un revenu annuel moyen compris entre 60 000 et 75 000 dollars, une superficie de logement moyen de 230 mètres carrés, une utilisation quotidienne d'électroménagers tels que le sèche-linge, et une consommation de lumière deux fois plus importante que celle du premier cluster. Les conclusions montrent ainsi une distinction claire entre les ménages à faible revenu avec des besoins essentiels et ceux à revenu élevé avec des équipements de loisir.

**Kernel K-Means** a identifié un sous-groupe spécifique de ménages très riches dans le second cluster, réduisant ainsi le nombre total d'échantillons dans ce cluster comparé à K-Means. Ce sous-groupe se distingue par des ménages possédant un logement encore plus grand, avec une superficie moyenne de 260 mètres carrés, et une utilisation accrue de tous les équipements. Cela signifie que Kernel K-Means a pu identifier un groupe de ménages extrêmement aisés, dont les habitudes de consommation énergétique sont encore plus marquées par des équipements de loisir et des comportements de consommation luxueux.

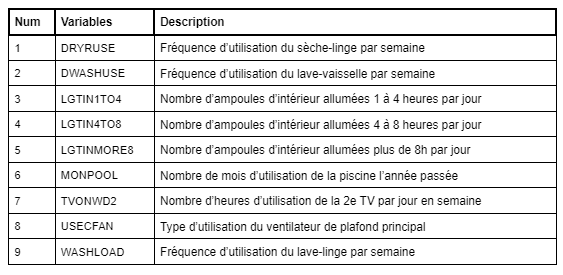
Les différences entre les deux clusters mettent en lumière les divers aspects de la consommation énergétique des ménages. Pour le premier cluster, la consommation d'énergie est principalement due à des besoins essentiels, reflétant des ménages vivant dans des logements modestes avec des revenus plus faibles. Les stratégies de réduction de la consommation d'énergie dans ce cluster pourraient se concentrer sur l'efficacité énergétique des appareils électroménagers et des pratiques quotidiennes. Pour le second cluster, la consommation d'énergie élevée est majoritairement attribuée à des équipements de loisir tels que la piscine ou la box satellite, caractéristiques de ménages à revenu élevé vivant dans de grands logements. Les efforts pour réduire la consommation d'énergie dans ce groupe pourraient inclure des mesures pour rendre ces équipements plus efficaces ou promouvoir des alternatives moins énergivores.

Pour compléter cette analyse, nous avons également appliqué l'algorithme DBSCAN, qui a également identifié deux clusters distincts. Cependant, il convient de noter que la majorité des points de données ont été classés comme des anomalies. Contrairement aux résultats des algorithmes précédents, les deux clusters détectés par DBSCAN regroupent principalement des ménages avec des revenus plus faibles. Bien que ces ménages présentent des différences en termes de revenus et de taille de logement, ils montrent des comportements similaires en termes de faible utilisation des équipements et appareils domestiques. Une distinction notable est que l'un de ces clusters ne possède pas de télévision du tout.

En somme, cette analyse révèle que les comportements de consommation énergétique diffèrent grandement selon le revenu et la taille du logement, influençant directement les stratégies potentielles d'économie d'énergie à adopter pour chaque groupe. Ces résultats mettent en lumière l'importance de cibler des mesures spécifiques d'efficacité énergétique adaptées aux caractéristiques et aux besoins des différents segments de la population, afin de maximiser les économies d'énergie et d'améliorer la durabilité environnementale.

## **3.5 Analyse descriptive des variables comportementales sélectionnées**

Après avoir segmenté les ménages en deux clusters distincts et analysé les caractéristiques de consommation énergétique, nous avons focalisé notre attention sur les variables comportementales parmi celles qui ont été sélectionnées. Rappelons que l'objectif initial de notre étude était d'analyser l'influence du comportement sur la consommation énergétique. Parmi les 17 variables issues de la sélection des variables, 9 sont spécifiquement liées aux comportements des ménages. En isolant ces variables comportementales, nous pouvons examiner plus en détail comment les habitudes et les pratiques des ménages impactent leur consommation d'énergie. Cette étape est cruciale pour comprendre les facteurs comportementaux spécifiques qui contribuent à la variation de la consommation d'énergie entre différents ménages.



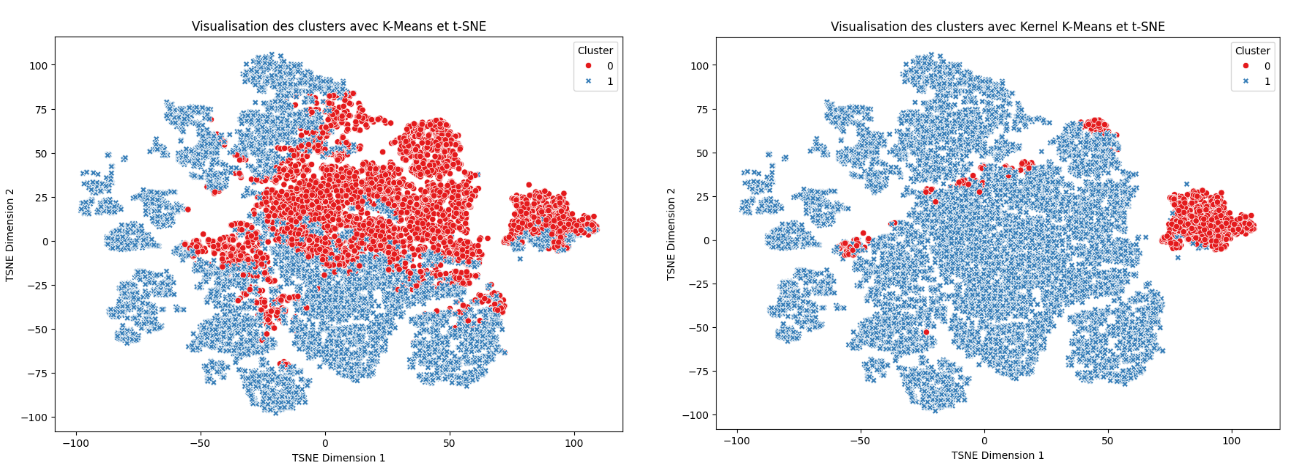
*Figure 9 : Tableau listant les variables comportementales*

### 3.5.1 Détermination du nombre optimal de clusters

Comme précédemment, nous avons déterminé le nombre optimal de clusters en utilisant plusieurs méthodes d’évaluation. La méthode du coude et le critère de Calinski-Harabasz ont toutes deux indiqué que le nombre optimal de clusters est 2. En revanche, le silhouette score a suggéré 7 clusters, tandis que le Davies-Bouldin index a recommandé 13 clusters. Ces divergences soulignent l’importance de considérer plusieurs critères pour une évaluation robuste du nombre de clusters, car chaque méthode utilise des critères différents et peut interpréter les données de manière distincte.

### 3.5.2 Deux clusters

L'analyse a révélé deux clusters principaux, comme observé précédemment. Les résultats obtenus avec les méthodes K-means et Kernel K-means montrent une similarité avec les clusters identifiés dans les analyses antérieures. Les deux méthodes ont confirmé la séparation entre les ménages à revenu faible et élevé. Kernel K-means met une nouvelle fois en évidence un sous-groupe spécifique parmi les ménages les plus aisés.

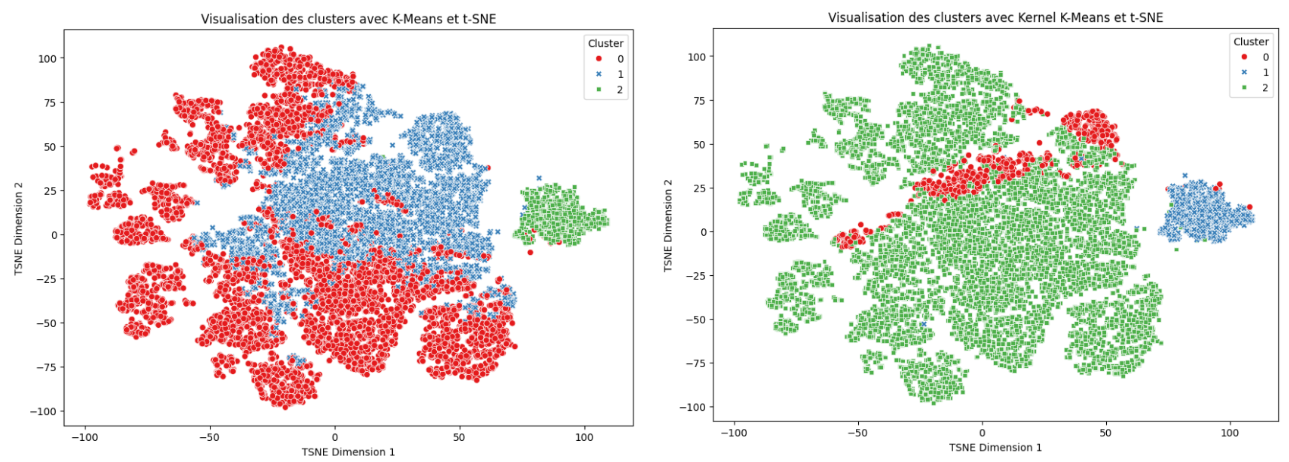


*Figure 10 : Visualisation du clustering des variables comportementales en 2 clusters*

### 3.5.3 Trois clusters

L'analyse avec trois clusters révèle des distinctions notables dans les comportements de consommation énergétique des ménages. Ce clustering met en lumière l'impact prépondérant de l'utilisation de la piscine sur la consommation d'énergie.

Le cluster le plus énergivore se caractérise par une utilisation élevée de la piscine, ce qui en fait le facteur principal influençant la consommation d'énergie dans ce groupe. Même dans les clusters où l'utilisation de la piscine est moins fréquente, cette variable reste fortement corrélée avec la consommation énergétique. Cette observation indique que, quel que soit le niveau d'utilisation de la piscine, elle exerce une influence significative sur les comportements énergétiques des ménages, soulignant l'importance de cet équipement dans les analyses de consommation d'énergie.



### 3.5.3 Six clusters

En appliquant les algorithmes de K-means et de Kernel K-means pour segmenter les données en 6 clusters, nous avons approfondi l'analyse des comportements de consommation énergétique des ménages. Les résultats montrent que les deux méthodes de clustering produisent des profils de segmentation similaires.

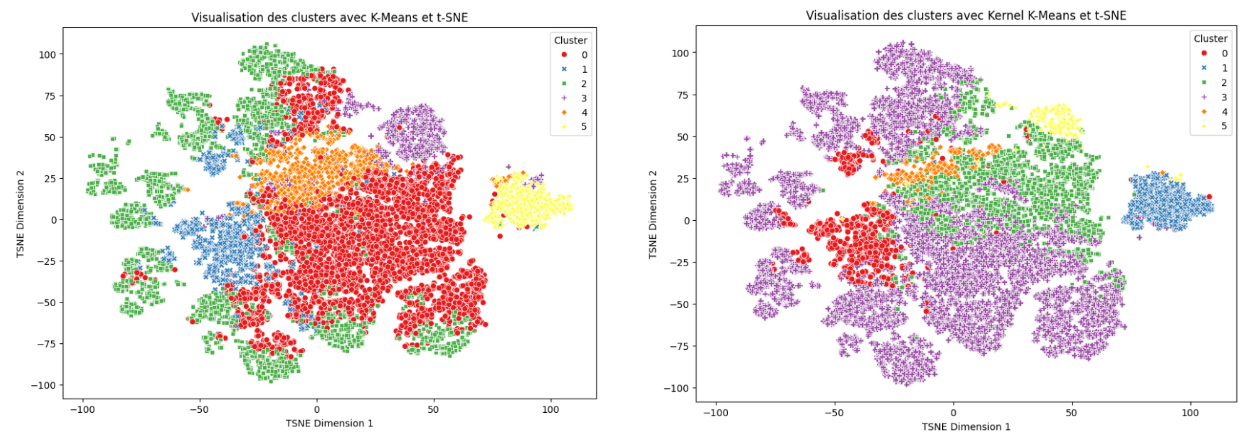
L'utilisation de la piscine se confirme comme le facteur le plus déterminant pour la consommation d'énergie, exerçant une influence prépondérante sur la consommation globale dans tous les clusters. En particulier, les ménages aisés, dont une proportion notable réside en Floride, tendent à utiliser leurs piscines presque la moitié de l'année, ce qui entraîne une consommation énergétique élevée. Ces ménages disposent souvent de grandes maisons, ce qui pourrait expliquer l'intensité de leur utilisation des piscines et des lumières, ou bien ils sont simplement moins sensibles à l'impact énergétique de ces équipements.

Les familles avec enfants se caractérisent par une utilisation importante des appareils de cuisine et de l'électroménager, utilisant des machines comme le lave-linge et le sèche-linge jusqu'à presque deux fois par jour pour chaque appareil. Elles possèdent également plusieurs télévisions, dont une est souvent utilisée pour les jeux vidéo des enfants, ce qui contribue à une consommation énergétique élevée.

Pour les ménages plus modestes, l'électroménager est le principal facteur de consommation, suivi par la télévision. Les familles vivant dans de petits logements montrent des comportements énergétiques où l'usage des appareils électroménagers et de la télévision est prédominant.

Chez les retraités, la télévision est le principal facteur de consommation d'énergie, avec une utilisation qui dépasse les 10 heures par jour. Ce comportement reflète leur tendance à passer une grande partie de la journée devant l'écran.

Cette analyse révèle ainsi des variations significatives dans les comportements énergétiques en fonction des ressources, des priorités des ménages, et des contextes géographiques.



*Figure 11 : Visualisation du clustering des variables comportementales en 6 clusters*

### 3.5.4 Dix clusters

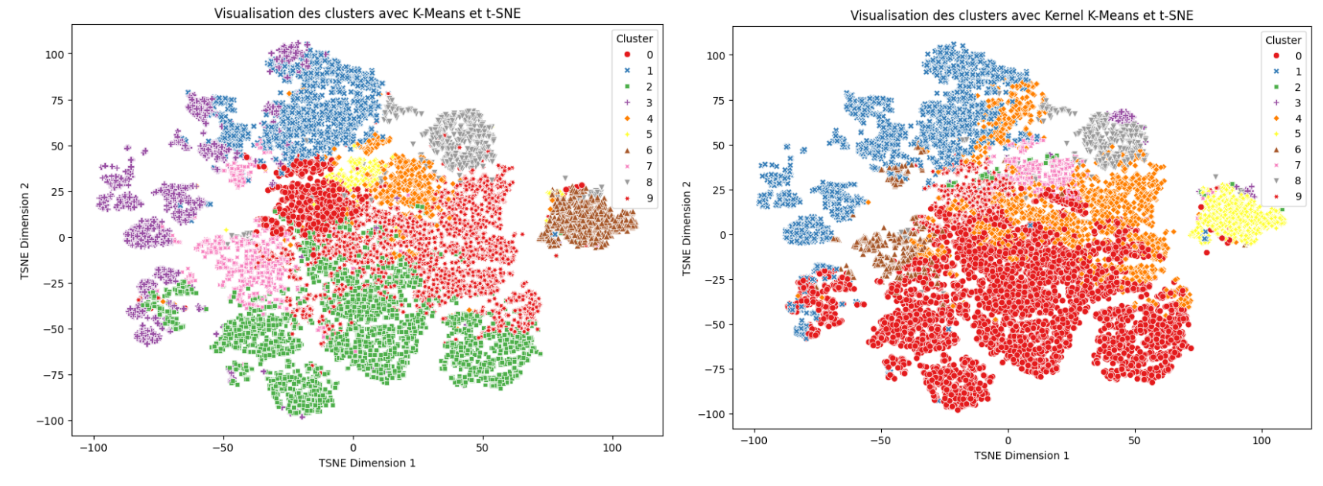
En appliquant les algorithmes de K-means et de Kernel K-means pour segmenter les données en 10 clusters, nous avons pu explorer plus en profondeur les subtilités des comportements énergétiques des ménages. Les deux méthodes ont mis en évidence plusieurs distinctions significatives, tout en confirmant certaines tendances observées dans les analyses précédentes.

L'utilisation de la piscine reste le facteur le plus énergivore, se démarquant clairement dans tous les clusters. Cependant, les autres comportements énergétiques se répartissent de manière plus fine dans ce clustering à 10 clusters.

La distinction notable concerne l'utilisation des lumières. Ce clustering différencie les ménages en fonction de la durée d'utilisation des lumières. Par exemple, l'utilisation intensive des lumières pendant une courte période (moins de 8 heures par jour) dans le cluster 7 s'est avérée plus énergivore que l'utilisation prolongée des lumières sur une plus longue durée (plus de 8 heures par jour) dans le cluster 9. Avec une recherche plus approfondie, cela peut s’expliquer, dans le cas du cluster 7, par le fait que 37,55% des ménages n'utilisent pas du tout d’ampoules LED, par rapport à seulement 9% dans le cluster 9. De plus, 72,27% des ménages du cluster 9 utilisent majoritairement des ampoules LED, contre seulement 15,92% dans le cluster 7. La durée d’utilisation des lumières n’est donc peut-être pas le facteur déterminant mais plutôt le type des ampoules.

Enfin, la télévision, comme dans les analyses précédentes, est ici reléguée au bas du classement des comportements énergétiques, en partie grâce à l'adoption généralisée des téléviseurs à faible consommation d'énergie, comme les téléviseurs à LED. Cependant, environ 10% des ménages possèdent encore des télévisions à forte consommation énergétique, telles que les téléviseurs à plasma ou à tube cathodique, qui ont un impact plus important sur la consommation énergétique globale par rapport aux autres appareils et usages domestiques.

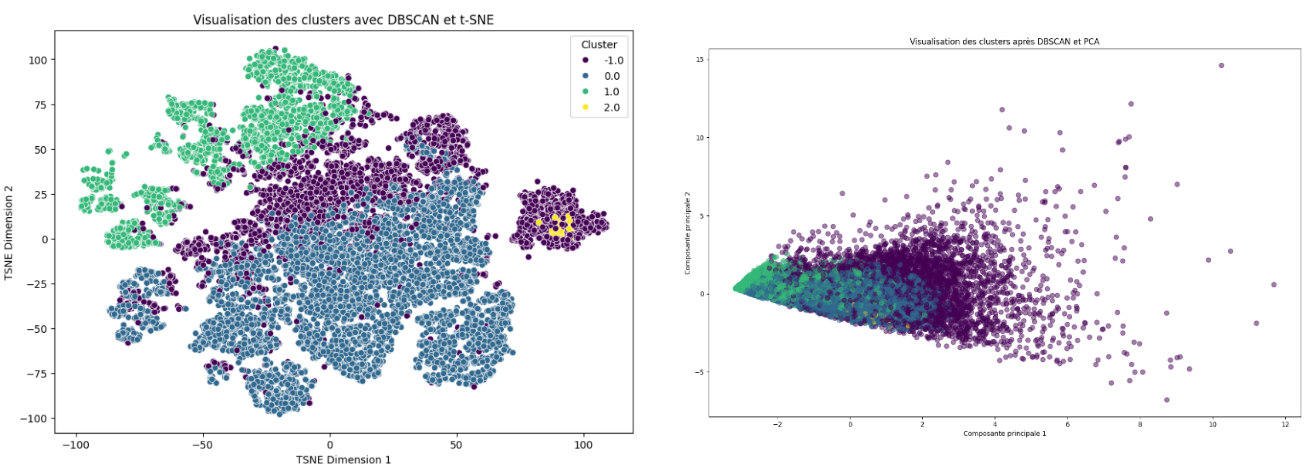
En segmentant les données en 10 clusters à l'aide des algorithmes de K-means et Kernel K-means, l'analyse a révélé des distinctions subtiles dans les comportements énergétiques des ménages. Cette segmentation a notamment mis en lumière des variations marquées dans l'utilisation des lumières, où une utilisation intensive sur une courte durée avec des ampoules non-LED s'est avérée plus énergivore qu'une utilisation prolongée avec des ampoules LED.



*Figure 12 : Visualisation du clustering des variables comportementales en 10 clusters*

### 3.5.5 DBSCAN

Le clustering effectué avec l'algorithme DBSCAN, en utilisant les paramètres eps=1 et min\_samples=18, a permis de définir deux clusters distincts de ménages sans piscine, dont l'un se caractérise également par l'absence de ventilateurs de plafond. Cette segmentation met en lumière des groupes de ménages avec des comportements énergétiques particuliers, notamment en l'absence d'équipements habituellement gourmands en énergie.



*Figure 13 : Visualisation du clustering des variables comportementales avec DBSCAN*

# **4. Discussion**

Dans cette étude, nous avons identifié plusieurs comportements clés influençant la consommation énergétique résidentielle. Cette section met en lumière les résultats principaux en lien avec des données et tendances pertinentes :

* Utilisation de la piscine

L'analyse a révélé que l'utilisation des piscines est le facteur le plus énergivore. Cette observation est renforcée par des données du marché [9] : entre 2019 et 2021, la vente de pompes à chaleur pour piscines a augmenté de 101%. Aux États-Unis, le marché des piscines résidentielles est significatif, avec 10,7 millions de piscines, dont 10,4 millions sont résidentielles, et près de 59% sont enterrées. En effet, les États-Unis sont le pays avec le plus de piscines privées au monde. Ces chiffres soulignent l'importance de l'efficacité énergétique des systèmes de chauffage des piscines. Des stratégies telles que l'optimisation des systèmes de chauffage ou l'utilisation de couvertures thermiques pourraient réduire considérablement à la fois la consommation énergétique associée aux piscines mais aussi l’évaporation de l’eau permettant ainsi une réduction de la consommation d’eau [28].

* Les appareils électroménagers

Les appareils électroménagers représentent également une part importante de la consommation d'énergie. Selon l’Agence Internationale de l’Énergie (AIE), les appareils modernes à haute efficacité peuvent réduire la consommation d'énergie jusqu'à 15% par rapport aux modèles plus anciens [10]. Les politiques encourageant l'achat d'appareils plus efficaces pourraient avoir un impact significatif sur la consommation globale d'énergie.

* Eclairage

L'éclairage prolongé, notamment les lumières extérieures laissées allumées la nuit, représente entre 10 et 15 % de la consommation d'électricité résidentielle moyenne [11]. L'adoption d'ampoules LED, qui consomment jusqu'à 10 fois moins d'énergie que les ampoules incandescentes et 6 à 8 fois moins que les ampoules halogènes [12], ainsi que la mise en place de systèmes de contrôle d'éclairage automatisés, peuvent entraîner des économies d'énergie significatives.

* Utilisation des appareils de cuisine

Même si les appareils de cuisine ne sont pas abordés dans cette étude, ils contribuent également à la consommation d'énergie domestique. L'adoption de pratiques de cuisson plus économes, telles que l'utilisation de fours à convection ou de cuisinières à induction, ainsi que l'optimisation de leur efficacité, peut contribuer à réduire cette consommation.

Cependant, l'impact relativement faible de la cuisine sur la consommation électrique des ménages américains s'explique par leurs habitudes culinaires. Les Américains passent en moyenne deux fois moins de temps à table que les Français (1h02 contre 2h13) [26], “les Américains ont de très belles cuisines, mais ne cuisinent pas” comme le dit [27], et préfèrent la malbouffe et les fast-foods, surtout dans les familles à revenu modeste. Les produits frais sont chers et moins accessibles que les produits industriels et les menus de fast food, ce qui limite les occasions de cuisiner [25].

“””Même si les appareils de cuisine ne sont pas abordés dans cette étude, ils contribuent également à la consommation d'énergie domestique. L'adoption de pratiques de cuisson plus économes, telles que l'utilisation de fours à convection ou de cuisinières à induction, ainsi que l'optimisation de leur efficacité, peut contribuer à réduire cette consommation.

Cependant, l'impact relativement faible de la cuisine sur la consommation électrique des ménages américains s'explique par leurs habitudes culinaires. Les Américains passent en moyenne deux fois moins de temps à table que les Français (1h02 contre 2h13), on dit aussi souvent que les Américains ont de très belles cuisines mais ils ne cuisinent pas, et préfèrent la malbouffe et les fast-foods, surtout dans les familles à revenu modeste. Les produits frais sont chers et moins accessibles que les produits industriels et les menus de fast food, ce qui limite les occasions de cuisiner.”””

* Télévision

Bien que l'impact de la télévision sur la consommation d'énergie soit relativement faible lorsqu'elle est utilisée de manière modérée, son effet cumulé peut devenir significatif lorsque son utilisation est généralisée. Ce phénomène est particulièrement notable chez les retraités, qui sont les plus grands consommateurs de télévision pendant la journée. Promouvoir des alternatives à la télévision, telles que des activités sociales ou physiques, pourrait non seulement réduire la consommation énergétique, mais aussi offrir des avantages pour leur santé et bien-être. Encourager ces comportements peut ainsi contribuer à des économies d'énergie tout en améliorant la qualité de vie des retraités.

* Les appareils intelligents

Bien que les appareils intelligents n’étaient pas largement présents dans cette étude en raison de leur faible consommation électrique individuelle et leur faible utilisation dans les ménages américains, ils figurent dans le dataset d'origine, notamment pour le contrôle de la température, de la télévision, des lumières et de la sécurité. Ces dispositifs possèdent un potentiel significatif pour réaliser des économies d'énergie à l'échelle globale. En tant qu'outils intelligents, ils sont capables de gérer et d'optimiser l'utilisation des autres appareils électroménagers, réduisant ainsi la consommation énergétique tout en améliorant la gestion des équipements domestiques. Leur adoption croissante pourrait ainsi jouer un rôle clé dans la transition vers des foyers plus économes en énergie et plus respectueux de l'environnement.

# **5. Conclusion**

Cette étude a réalisé une analyse approfondie de la consommation énergétique résidentielle en utilisant des méthodes de prétraitement, de sélection de caractéristiques et de clustering pour comprendre l'influence des comportements des ménages et identifier des groupes distincts de consommation énergétique.

Nous avons commencé par un prétraitement rigoureux des données, réduisant le nombre d'échantillons de 18 496 à 17 312 et les variables de 799 à 208. Quatre méthodes de sélection des caractéristiques ont permis de réduire le nombre de variables pertinentes à 17, dont 9 étaient comportementales.

L'application des algorithmes de clustering, tels que K-Means et Kernel K-Means, a révélé des variations significatives dans les comportements énergétiques des ménages. Les analyses ont permis d'identifier plusieurs comportements clés en fonction de leur impact énergétique. Parmi ces comportements, l'utilisation des piscines se distingue comme le facteur le plus énergivore. Les appareils électroménagers ont également un impact significatif, tandis que l'éclairage, en particulier lorsqu'il utilise des ampoules non-LED, contribue de manière notable à la consommation d'énergie. L'utilisation des appareils de cuisine et le visionnage de la télévision, bien que moins énergivores, contribuent également à la consommation globale.

Cette répartition souligne l'importance de mettre en place des stratégies adaptées pour gérer la consommation d'énergie de manière efficace. Les équipements de loisirs, en particulier les piscines, nécessitent des interventions ciblées pour améliorer leur efficacité énergétique. Les appareils électroménagers doivent être optimisés pour minimiser leur impact, tandis que des mesures telles que l'adoption d'ampoules LED et l'instauration de systèmes de contrôle automatisés peuvent offrir des économies notables en matière d'éclairage.

À l’avenir, les technologies numériques, telles que le streaming vidéo et le cloud computing, qui sont très gourmands en ressources, pourraient accroître significativement la consommation énergétique résidentielle. En 2023, le secteur informatique représente environ 10 % de la consommation mondiale d'électricité, en hausse par rapport aux 7 % en 2016, et cette proportion pourrait atteindre 20 % d'ici 2030. Le streaming vidéo, qui constitue environ 80 % du trafic web mondial, contribue largement à cette tendance [29]. L'impact croissant de ces technologies sur la consommation d'électricité nécessite une attention particulière pour élaborer des stratégies efficaces visant à atténuer leur effet sur la demande énergétique.

D’un autre côté, il est pertinent de noter que la prédiction de la consommation électrique pourrait bénéficier de l'intégration des technologies numériques émergentes. Ces technologies jouent un rôle de plus en plus crucial dans la gestion et la réduction de la consommation d'énergie résidentielle. L'Internet des objets (IoT) et les objets connectés permettent une collecte de données en temps réel sur la consommation énergétique, offrant des opportunités pour des analyses plus fines et des prévisions plus précises. Les systèmes de gestion de l'énergie domestique (HEMS) peuvent surveiller et optimiser l'utilisation de l'énergie, tandis que les réseaux intelligents (smart grids) permettent une distribution plus équilibrée et efficace de l'électricité. L'Internet du comportement (IoB) pourrait également jouer un rôle crucial en permettant une compréhension approfondie des habitudes et des comportements des utilisateurs. Malheureusement, pour cette recherche, nous n'avons pas pu accéder à des données aussi détaillées, ce qui limite notre capacité à tirer pleinement parti de ces technologies émergentes. Néanmoins, elles pourraient transformer la manière dont nous anticipons les besoins énergétiques et optimisons la gestion de la consommation, en offrant des outils puissants pour améliorer l'efficacité énergétique et réduire l'empreinte carbone des ménages à l'avenir.

“””Il serait intéressant de poursuivre cette étude en utilisant un ensemble de données plus complet, incluant des informations provenant d'objets connectés. Un tel dataset permettrait de mieux comprendre l'impact réel de ces technologies sur la réduction de la consommation énergétique. De plus, il serait pertinent de mener cette expérimentation dans des pays où l'adoption des objets connectés est plus répandue, comme la Chine. En effet, la Chine, avec son marché technologique avancé, offre des opportunités pour collecter des données plus précises et diversifiées sur l'utilisation des objets connectés au quotidien. Cela permettrait de mieux évaluer l'efficacité des solutions technologiques dans la gestion énergétique et d'affiner les stratégies d'optimisation de la consommation à grande échelle. En combinant ces nouvelles données avec des méthodes d'analyse approfondies, il serait possible de dégager des tendances plus robustes et de proposer des recommandations mieux adaptées aux contextes technologiques émergents.”””

Aujourd'hui, alors que nous atteignons le Jour du Dépassement de cette année 2024, cette étude souligne l'impact environnemental crucial de nos comportements énergétiques. Le Jour du Dépassement marque la date à laquelle l'humanité a consommé toutes les ressources que la Terre peut renouveler en une année. Adopter des stratégies ciblées pour optimiser la consommation d'énergie au niveau des ménages est essentiel pour favoriser une prise de conscience accrue et des pratiques plus durables, contribuant ainsi à réduire notre empreinte écologique et à préserver les ressources de la planète.

# **Références**

1. U.S. Energy Information Administration, 2020 Residential Energy Consumption Survey. <https://www.eia.gov/consumption/residential/data/2020/>
2. Évolution du nombre d'habitants sur Terre entre 1950 et 2024 et projections jusqu'en 2100. Statista. <https://fr.statista.com/statistiques/564933/population-mondiale-jusqu-en-2080/>
3. Population. Nations Unies. <https://www.un.org/fr/global-issues/population>
4. D. Chareyron, H. Horsin-Molinaro, B. Multon. Concepts et chiffres de l'énergie : la consommation de l’électricité domestique en France. Culture Sciences Physiques, ENS Lyon. <https://culturesciencesphysique.ens-lyon.fr/ressource/chiffres-energie-electricite-domestique.xml>
5. H. M. Sani, S. O. Tehrani, B. Behkamal, H. Amintoosi : Extracting Effective Features for Descriptive Analysis of Household Energy Consumption Using Smart Home Data. 2019.
6. M. Heinrich, M. Ruellan, L. Oukhellou, A. Samé, J.-P Lévy : From energy behaviors to lifestyles: Contribution of behavioral archetypes to the description of energy consumption patterns in the residential sector. 2022.
7. T. F. Sanquist, H. Orr, B. Shui, A. C. Bittner : Lifestyle factors in U.S. residential electricity consumption. 2012.
8. Analyse du marché : Quelle est la taille du marché mondial de la piscine ?. Aquark. <https://www.aquark.com/fr/analyse-du-marche-mondial-de-la-piscine/>
9. U.S. Energy Information Administration. A call to action on efficient and smart appliances. <https://www.iea.org/articles/a-call-to-action-on-efficient-and-smart-appliances>
10. Engie. LED, halogène, connectée… quelle est la consommation d’une ampoule ?. <https://particuliers.engie.fr/economies-energie/conseils-economies-energie/conseils-calcul-consommation/consommation-ampoule.html>
11. Élisabeth Chesnais. Ampoules LED De belles économies. UFC Que Choisir. <https://www.quechoisir.org/guide-d-achat-ampoules-basse-consommation-led-n11547/>
12. M. Labiadh. Methodology for construction of adaptive models for the simulation of energy consumption in buildings. 2022.
13. H. Elayan, M. Aloqaily, F. Karray, M. Guizani. Decentralized IoB for Influencing IoT-based Systems Behavior. 2024.
14. H. Elayan, M. Aloqaily, F. Karray, M. Guizani. Internet of Behavior and Explainable AI Systems for Influencing IoT Behavior. 2023.
15. J. Wang, X. Chen, F. Zhang, F. Chen, Y. Xin. Building Load Forecasting Using Deep Neural Network with Efficient Feature Fusion. 2021.
16. J. W. Chan, C. K. Yeo. Electrical Power Consumption Forecasting with Transformers. 2022.
17. D. Hadjout, J.F. Torres, A. Troncoso, A. Sebaa, F. Martínez-Alvarez. Electricity consumption forecasting based on ensemble deep learning. 2021.
18. D. Hadjout, A. Sebaa, J. F. Torres, F. Martínez-Álvarez. Electricity consumption forecasting with outliers handling based on clustering and deep learning with application to the Algerian market. 2023.
19. D. Syed, H. Abu-Rub, A. Ghrayeb, S. S. Refaat. Household-Level Energy Forecasting in Smart Buildings Using a Novel Hybrid Deep Learning Model, 2021.
20. M. Alhussein , K. Aurangzeb, S. I. Haider. Hybrid CNN-LSTM Model for Short-Term Individual Household Load Forecasting. 2020.
21. X. Li, Y. Zhong, W. Shang, X. Zhang, B. Shan, X. Wang. Total electricity consumption forecasting based on Transformer time series models. 2022.
22. Y. Liu, D. Zhang, H. B. Gooi. Optimization Strategy Based on Deep Reinforcement Learning for Home Energy Management. 2020.
23. S.-H. Kim, C. Lee, C.-H. Youn. An Accelerated Edge Cloud System for Energy Data Stream Processing Based on Adaptive Incremental Deep Learning Scheme. 2020.
24. EuropUSA. Etats-Unis : les habitudes alimentaires. <https://www.europusa.com/vivre-aux-etats-unis/vie-personnelle-aux-usa/vie-quotidienne/etats-unis-les-habitudes-alimentaires/>
25. SudOuest. Les Français passent deux fois plus de temps à table que les Américains. <https://www.sudouest.fr/premium/art-de-vivre/les-francais-passent-deux-fois-plus-de-temps-a-table-que-les-americains-3136883.php>
26. Christina Gierse. Les Français et la nourriture aux Etats-Unis : un sujet « touchy ». Studyrama. <https://www.studyrama.com/pro/destination/les-francais-et-la-nourriture-aux-etats-unis-un-sujet-touchy-21891.html>
27. Guide Piscine. <https://www.guide-piscine.fr/pro/marche-de-la-piscine/piscine-eco-responsable/l-impact-d-une-couverture-sur-la-consommation-en-eau-et-en-electricite-d-une-piscine-5620_A>
28. Christophe Magdelaine. Streaming : quelles émissions de CO2 et consommation d'énergie ?, 07/03/2024. https://www.notre-planete.info/actualites/247-streaming-Internet-electricite-CO2

Abstract. Household energy consumption represents a significant share of overall energy consumption. To deepen our understanding of surplus energy production, management, and storage, it is crucial to elucidate the implicit patterns of consumer behaviors and identify factors influencing their performance. This paper aims to analyze the household energy consumption pattern, with a focus on the impact of user behavior on energy consumption. To this end, we focus on selecting the most relevant feature subset from a large dataset, which allows for better insights, reduced computation time, and improved predictive performance. The results of this study can help policymakers examine household behaviors at various levels of society to ensure that their quality of life is adequate or needs improvement.

Keywords: Feature selection · Smart home · Household energy consumption · Descriptive analysis - User behavior - Residential energy efficiency

1 Introduction

The world population reached 8.2 billion in 2024 [1], one billion more than

in 2010, and is expected to cross the 10 billion mark before the end of the century

according to United Nations forecasts [2,3].This rapid growth, combined with urbanization and increasing energy needs, is contributing to a significant increase in household energy consumption, which represents about 25% of global energy consumption [3]. In metropolitan France, households currently account for 36% of electricity

consumption, making this sector the most energy-intensive in proportion, ahead

of the tertiary and industrial sectors [4]. This trend raises significant concerns about energy sustainability, particularly in terms of the increasing strain on the electricity grid, rising carbon emissions, the challenges of transitioning to renewable energy sources, and the need to balance electricity production with growing demand.

The rise of smart Internet of Things (IoT) devices, whose number is expected to reach 50 billion units by 2025 and 100 billion by 2030 [2], opens up new perspectives for monitoring and optimizing energy consumption. These devices collect massive amounts of data on user habits, preferences and behaviors, offering an unprecedented opportunity to understand

and manage energy consumption more efficiently. The Internet of Behavior (IoB) is a new paradigm aimed at analyzing data related to user behavioral trends to achieve specific objectives [1,2,3]. This data, collected by smart devices, reflects user behavior, habits, and lifestyle across various domains such as health [4], transportation [5,6], education [7,8], and energy consumption [9,10]. IoB enables the development of energy management models within residential spaces by observing how occupants' behavior influences energy consumption. By continuously monitoring user behavior, IoB provides valuable insights into how energy consumption can be optimized. However, the challenge remains: how to interpret these data to develop efficient energy management models in residential spaces, based on occupant behavior? While existing approaches mainly focus on selecting features related to the devices used [Lifestyle factors, From energy] or the homes

themselves (surface area, number of bedrooms) [Mouna labiadh], our approach is distinguished by a focus on behavioral features. The way household members use their appliances, such as the frequency of use and the mode of operation, can greatly influence overall energy consumption. For example, excessive use of appliances like washing machines, dryers, or air conditioning systems can lead to significantly higher energy bills. Understanding these usage patterns enables us to identify opportunities for optimization, such as reducing the frequency of use, switching to energy-saving modes. We seek to understand how human behaviors influence energy consumption, thus offering a more detailed analysis and an opportunity to improve overall energy efficiency.

In this paper, we analyze the RECS2020 dataset, which contains information on residential energy consumption of US households in 2020. We develop learning models to identify correlations between user behaviors and their energy expenditures in residential spaces. Our methodology includes collecting and preprocessing behavioral data, selecting relevant features, and applying clustering algorithms to segment households into meaningful groups. These clusters are then analyzed to extract information on consumption habits.

The remainder of this article is structured as follows: In Section 2, we review the related work in the field, providing context for our study. Section 3 presents the problem definition, outlining the challenges and objectives of our research. Section 4 details the methodology, including data preprocessing, clustering techniques, correlation analysis, and data exploration. In Section 5, we discuss the result analysis, where we examine the findings and their implications. Finally, Section 6 concludes the article, summarizing the key insights and potential avenues for future research.

2 Related Work

In this section, we review related work in three main areas: energy consumption

prediction, feature extraction methods, and techniques for influencing energy use

behaviors

Prediction models are essential tools in understanding and forecasting energy consumption patterns. These models leverage historical data and various influencing factors to estimate future energy usage, enabling more efficient energy management. Mouna Labiadh [13] presents a novel methodology for energy consumption prediction in buildings, especially where historical data is lacking. This approach utilizes a Siamese MLP model combined with KNN and various temporal models like SVR, MLP, LSTM, CNN, and Seq2seq, focusing primarily on the physical characteristics of buildings. Wang et al. [16] developed a deep

convolutional neural network (DCNN) based on ResNet, integrating temporal

and meteorological features for accurate electricity consumption forecasting. Chan and Yeo [17] employed a Sparse Transformer, which provides comparable accuracy to RNNs while being faster. Hadjout et al. [18,19] used a combination of LSTM, GRU, TCN, and

SARIMA models, enhanced by Ensemble Learning techniques for electricity consumption prediction. Syed et al. [20] explored various LSTM architectures, such as bidirectional and unidirectional layers, to forecast electricity consumption. Alhussein et al. [21] integrated hybrid architectures that combine CNN and LSTM for better forecasting accuracy. Li et al. [22] developed the Trans-T2V model, combining a Transformer with Time2Vec for refined electricity consumption forecasts.

Several studies have focused on identifying key factors that influence energy consumption. For instance, Sani et al. [5] analyzed U.S. household energy consumption using the RECS 2015 dataset, but primarily focused on socio-economic factors rather than behavioral aspects. Sanquist et al. [7] identified lifestyle-related factors such as air conditioning usage, laundry practices, and personal computer usage, based on RECS data from 2001 and 2005. Like our study, Heinrich et al. [6] aimed to build behavioral archetypes to better understand energy consumption habits in the residential sector. The data, drawn from the ENERGIHAB project of the French National Research Agency, included 35 variables covering hygiene, food, heating, lighting, leisure practices, and housing occupation from 1363 households in Île-deFrance. This study successfully identified seven distinct behavioral archetypes.

Recent advancements have explored ways to influence and modify energy

consumption behaviors. Haya Elayan et al. [14] proposed a decentralized IoB framework aimed at predicting electricity consumption and

influencing IoT device behavior. This framework, using French household electricity consumption data and LSTM models, shows promise in reducing energy

consumption by directly affecting the operation of connected devices. Although

this framework has been tested on a limited number of devices, it offers an optimistic perspective on the possibility of reducing electricity consumption not only

at the household level, but also at that of an entire population. Another study

by Elayan et al. [15] combined IoB with explainable AI (XAI) to optimize energy

use decisions while making these decisions transparent to users. The integration

of an Energy Monitor & Controller (EMC) and an "Explainer" component helps

users understand and accept energy-saving recommendations, encouraging more

sustainable behavior. A well-explained decision, which clearly shows the benefits

in terms of energy savings and cost reduction, has the potential to motivate users

to adopt more sustainable practices.

The existing literature offers a variety of advanced techniques for

energy consumption prediction, feature extraction, and influencing the behaviors

of connected devices. Predictive models, such as those based on LSTM [Mouna Labaidh, Building Load, Electrical power, Hadjout, Household level] or other

neural networks [Building Load, Electrical power, Household level, Hybrid ], are already well developed to anticipate energy needs. Similarly,

sophisticated feature extraction methods allow to analyze socio-demographic factors and building characteristics influencing consumption. Influence techniques

also show promising potential to modify user behaviors and optimize energy

use. However, this study stands out by focusing specifically on the identification and segmentation of household behaviors, an aspect often underexplored in

previous works.Many prior studies were limited by geographic scope, a restricted number of variables, or biases in data collection. Some relied on outdated data, had missing information, or did not focus on user behavior, which constrained their ability to provide a comprehensive analysis of energy consumption patterns. By combining these existing techniques with in-depth behavioral analysis, we are planning to improve the effectiveness of energy interventions.

Our approach aims to reduce energy consumption more effectively by targeting specific daily habits and adapting solutions to both occupant behavior and lifestyle patterns. This originality allows

us to better understand how behaviors can be modified to reduce the energy

footprint, thus contributing to more efficient and targeted energy policies.

3 Problem definition

Managing household energy consumption is a major challenge in the current

context of climate change and the transition to renewable energy sources. A

better understanding of household energy consumption behaviors can help identify levers for action to reduce energy demand, improve energy efficiency, and

decrease greenhouse gas emissions.

The problem lies at the intersection of several issues: environmental, economic, and social. Household energy consumption contributes significantly to

greenhouse gas emissions, which has a direct impact on climate change. Reducing energy consumption is essential to achieve the emission reduction targets set

by various international agreements. On the economic level, energy represents a

significant share of household expenditure. By identifying energy-intensive behaviors and proposing solutions to reduce them, it is possible to reduce household

energy bills, which can have a positive impact on purchasing power. On the social

level, there are significant disparities in energy consumption based on income,

housing size and household equipment. Understanding these disparities is crucial

to propose equitable and adapted measures for each type of household.

The main objective of this study is to analyze the influence of behavior on energy consumption. This involves examining how various behaviors impact energy use, identifying distinct consumption patterns, and understanding the key factors driving these patterns. By conducting a thorough analysis of household data, including descriptive statistics and clustering, we aim to uncover the relationship between behavior and energy consumption. The goal is to provide insights into how different behaviors contribute to energy use, which allow us to offer targeted recommendations for reducing consumption based on these findings.

4 Methodology

The objective of this study is to highlight the user behaviors that have the greatest

impact on household energy consumption, based on the 2020 RECS dataset. We

have selected three target variables: the total amount of electricity used

in kilowatt-hours (kWh), the total amount of electricity used in BTU (BTUEL),

and the total cost of electricity consumed in dollars (DOLLAREL).

In this section, we first describe the dataset used. Then, we detail the

different preprocessings applied to prepare the data for analysis, as well as the feature selection methods employed to identify the most significant behaviors that impact energy consumption. Finally, we analyze the results obtained after the application of clustering algorithms and studying the correlations that allowed us to identify the characteristics that have the greatest influence on the energy consumed.

4.1 Dataset Overview

The Residential Energy Consumption Survey (RECS) is a comprehensive survey

conducted in the United States to collect information on the residential energy

consumption of American households. Conducted by the Energy Information

Administration (EIA), this survey is one of the most comprehensive and reliable

sources of data on energy use in the residential sector. [1]

The 2020 edition of the RECS collects data on various aspects of household

energy consumption, including the types and quantities of fuels used such as electricity, natural gas, propane, and fuel oil, dwelling characteristics, household appliances, heating and cooling systems, as well as

recent behaviors such as the use of solar energy or the location of electric vehicle

charging. The survey thus covers a wide range of 799 variables, allowing for an

in-depth and multidimensional analysis of energy consumption habits.

The RECS uses a representative sample of 18496 American households to ensure that the results can be generalized to the entire population. Detailed surveys

are sent to selected households and can be administered as paper questionnaires,

online, by telephone interviews or during on-site visits.

4.2 Preprocessing

To ensure the quality and relevance of the analyses, several preprocessing steps

were applied to the RECS 2020 data. These steps are crucial to clean and prepare

the data before applying feature selection techniques and clustering algorithms.

**Removal of variables concerning energies other than electricity** Since

our research aims to identify user behaviors influencing electricity consumption, we

removed 117 variables related to other types of energy, such as natural gas,

propane, fuel oil and wood. By excluding these variables, we were able

to focus our analysis specifically on the factors influencing electricity consumption. This approach reduces the complexity of the dataset and facilitates the

interpretation of the results. After this removal, we retained a total of 682 variables

**Removal of imputation and calibration indicator variables** Imputation indicator variables show whether the values for other variables in the dataset have been imputed, meaning they represent estimates rather than actual data. Calibration variables adjust the weights of responses to ensure the sample accurately reflects the broader population. To refine our analysis and ensure data relevance, we removed 407 of these

variables used by the EIA. While these variables are crucial to the EIA’s internal processes, they are not necessary for our study, which focuses specifically on electricity consumption behaviors. After these removals, 275 variables remain for our analysis

.

**Encoding categorical variables** To prepare our dataset for further analysis, we converted the categorical variables into a form that machine learning algorithms can understand. We used the LabelEncoder method from the scikit-learn library. This technique transforms each categorical value into a unique numeric value, indicating the category it belongs to. In our dataset, we identified 7 categorical variables that required encoding. Applying the LabelEncoder to these variables allowed us to convert the categorical data into a numeric representation while preserving the integrity and meaning of the original information.

**Missing Values Management** To ensure the quality and completeness of our

dataset, we adopted a pragmatic approach to handling missing values. The first step was to identify all rows in the dataset that contained one or more missing values. We identified 290 samples containing a NaN value , which represents approximately 1.57% of the dataset. Since this proportion is very small, we decided to remove these samples to simplify data processing. After removing rows with missing values, the number of households in our dataset decreased from 18,496 to 18,206. This approach ensures that our analysis is based on a complete dataset, thereby m

**Reduction of target variables** At the beginning of our study, we identified

three main target variables related to electricity consumption: the total amount

of electricity used in kilowatt hours (kWh), the total amount of electricity used

in BTUs (BTUEL), and the total cost of electricity consumed in dollars (DOLLAREL). However, after performing a Pearson correlation analysis between these variables, we found that some of them were highly correlated. In particular, the BTUEL variable was highly correlated with kWh, because these two measures are interconvertible (1 kWh is equivalent to approximately 3,412 BTUs). As a result, we decided to reduce the number of target variables to two: kWh and DOLLAREL. This reduction simplifies the analysis while retaining

the essential measures to assess electricity consumption. Following this reduction, we have added these 2 variables to our dataset total, bringing the number of variables to 277.

Supprimer legende de l’image

**Reducing the number of variables** As part of data preprocessing, we performed a Pearson correlation analysis to identify highly correlated variable pairs in our dataset and reduce redundancy. This analysis led to the removal of 69 variables. For example, we removed the variable STATE\_POSTAL and kept STATE\_NAME to avoid redundancy between state zip codes and their names. Similarly, we removed the variable TELLWORK, which indicates that a household member teleworks, in favor of more precise variables such as TELLDAYS, TLDESKTOP or TLLAPTOP, which provide more information on telework arrangements. This approach was also applied to other domains, such as energy assistance (PAYHELP), electric vehicles (ELECVEH) or housing type (STUDIO), where we chose to keep the variables providing the most detailed and relevant information. The correlation matrix below illustrates the highly correlated variables and supports our decisions for variable reduction by showing how some variables are closely related. After this reduction process, 69 variables were removed, leaving a total of 208 variables for further analysis.

Supprimer titre de l’image

**Reducing the number of samples** Isolated data, often considered as outliers,

can bias machine learning models and distort the results of statistical analyses.

It is therefore essential to identify and remove them to obtain more accurate and

reliable analyses, thus allowing a better understanding of the data.

In our study, we used the Isolation Forest algorithm [mettre ref] to identify these isolated

data. Isolation Forest isolates anomalies by building isolation trees. Each tree

randomly partitions the data until each point is isolated in a leaf. Anomalies,

being rare and different, are isolated more quickly, meaning that they require

fewer splits to be separated from other points. Each data point is assigned an

anomaly score based on the average depth at which it is isolated in the trees. The

shallower the depth, the more likely the point is to be an anomaly. This method

detected 894 samples considered as outliers, representing approximately 5% of the dataset. Removing these points reduced the total number of households in

our dataset from 18,206 to 17,312. By removing these isolated data points, we

significantly improved the quality of the dataset, allowing for more accurate and

robust analyses.

Modifier image

**Preprocessing Conclusion** Following rigorous preprocessing steps, we significantly refined our dataset to make it more relevant and usable for our analysis.

The number of samples was reduced from 18,496 to 17,312, following the

removal of 894 outliers identified as isolated and 290 samples containing at least

one missing value. This reduction allowed us to maintain a more coherent and

representative database.

Regarding variables, we simplified our dataset by reducing the total number

of variables from 799 to 208. This reduction was achieved by eliminating variables

related to types of energy other than electricity, imputation and calibration indicator variables,

as well as those that were highly correlated. This approach reduced redundancies

and focused our analysis on the most relevant factors.

These preprocessing steps not only optimized data quality, but also facilitated more precise and meaningful analyses. Ultimately, these adjustments help

to obtain more reliable results and deepen our understanding of electricity consumption behaviors.

| pre-processing steps | Number of variables | Number of samples |
| --- | --- | --- |
| original dataset | 18496 | 799 |
| Removal of variables concerning energies other than electricity | 18496 | 682 |
| Removal of imputation and calibration indicator variables | 18496 | 275 |
| Encoding categorical variables | 18496 | 275 |
| Missing Values Management | 18206 | 275 |
| Reduction of target variables | 18206 | 277 |
| Reducing the number of variables | 18206 | 208 |
| Reducing the number of samples | 17312 | 208 |
| Conclusion | 17312 | 208 |

4.3 Variable selection

Variable selection is a crucial step in data preprocessing, aiming to improve the

quality of predictive models by identifying the most relevant variables. In our

study, we applied several methods to refine our choice of variables.

**Correlation across the dataset** The first method of variable selection is based

on the analysis of the correlation between all the variables in the dataset and

the target variables. We first calculated the correlation between each variable

and the target variable KWH to identify those with a significant relationship

with electricity consumption. The variables whose correlation with KWH was

higher than the average of the calculated correlations were selected. Similarly, we

evaluated the correlation between each variable and the second target variable,

DOLLAREL, to identify the variables that contribute the most to the cost of electricity consumption. The variables whose correlation with DOLLAREL was also higher than the

average of the calculated correlations were retained. Finally, the relevant variables for each of the two targets were crossed to obtain their intersection, which

made it possible to select 76 important variables for the two target variables,

guaranteeing their relevance in both analysis contexts.

**Information Gain on the entire data set** The DecisionTreeRegressor is a

regression algorithm that builds a decision tree to predict continuous values,

such as energy consumption (KWH) or cost (DOLLAREL).

The algorithm assesses the importance of

each feature by measuring the reduction in impurity, quantified by the squared

error, brought by each division of the tree. At each node, it selects the feature

that minimizes this impurity, thus optimizing the predictions.

After determining the importance of each feature, we rank these features based on their contribution to reducing the overall squared error. A cumulative curve is then plotted to visualize the contribution of each feature.

Modifier le tableau

Fusionner les tables des méthodes 2 et 4

[1] <https://www.ined.fr/fr/tout-savoir-population/memos-demo/focus/2024-les-nations-unies-publient-de-nouvelles-projections-de-population-mondiale>

[2] <https://www.telecom-sudparis.eu/actualite/objets-connectes-50-milliards-demetteurs-de-co2/>

[3] María del P. Pablo-Romeroa, Rafael Pozo-Barajasb, and Rocío Yñiguez. GLOBAL CHANGES IN RESIDENTIAL ENERGY CONSUMPTION. JEL code: Q2, Q48, Q58, R1, R58

[1] Sun, Jiayi, Gan, Wensheng, Chao, Han-Chieh, Philip, S Yu, Ding, Weiping. Internet of behaviors: A survey. IEEE

Internet of Things, vol. 10, no. 13, pp. 11117 – 11134, 2023.

[2] Zhao, Q., Li, G., Cai, J., Zhou, M., and Feng, L. A Tutorial on Internet of Behaviors: Concept, Architecture,

Technology, Applications, and Challenges. IEEE Communications Surveys & Tutorials, vol. 25, no. 2, pp. 1227 –1260,

2023.

[3] Ziani, L., Khanouche, M. E., Belaid A.: Internet of behaviors: A literature review of an emerging technology. In 1

st

Int. Conf. on Big Data, IoT, Web Intelligence and Applications, pp. 42–47, Sidi Bel Abbes, Algeria, 2022,

[4] Javaid, M., Haleem, A., Singh, R.P., Khan, S., Suman, R.: An extensive study on internet of behavior (iob) enabled

healthcare-systems: Features, facilitators, and challenges. BenchCouncil Transactions on Benchmarks, Standards

and Evaluations, 100085, 2023.

[5] Mezair, Tinhinane, Djenouri, Youcef, Belhadi, Asma, Srivastava, Gautam, Lin, Jerry Chun-Wei. Towards an

Advanced Deep Learning for the Internet of Behaviors: Application to Connected Vehicles. ACM Transactions on

Sensor Networks, vol. 19, no. 2, pp. 1–18, 2022.

[6] Song, Qun, Tan, Rui, Wang, Jianping. Towards Efficient Personalized Driver Behavior Modeling with Machine

Unlearning. In Proceedings of Cyber-Physical Systems and Internet of Things, pp. 31–36, 2023.

[7] Embarak, O.H.: Internet of behaviour (IoB)-based AI models for personalized smart education systems. Procedia

Computer Science 203, 103–110, 2022.

[8] Embarak, Ossama. An adaptive paradigm for smart education systems in smart cities using the internet of

behaviour (IoB) and explainable artificial intelligence (XAI). 8th Int. Conference on Information Technology Trends,

pp. 74–79, 2022.

[9] Elayan, H., Aloqaily, M., Karray, F., Guizani M.: Internet of behavior (IoB) and explainable AI systems for

influencing Internet of Things behavior. IEEE Network, 2022.

[10] Elayan, Haya, Aloqaily, Moayad, Karray, Fakhri, Guizani, Mohsen. Decentralized IoB for Influencing IoT-based

Systems Behavior. IEEE Int. Conference on Communications, pp. 3340–3345, 2022.