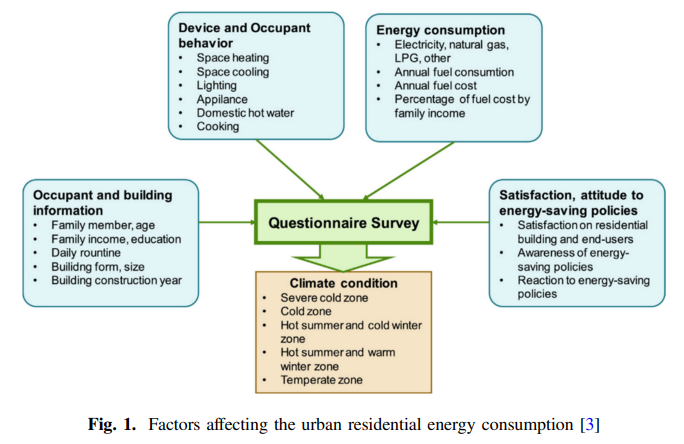
Synthèse : Extracting Effective Features for Descriptive Analysis of Household Energy Consumption Using Smart Home Data

Article de 2019

Identifier les facteurs influençant la consommation d’énergie dans le comportement des ménages

## **1 Introduction**

Facteurs influençant la consommation d’énergie urbaine d’après une étude en Chine



Mais il faut faire une sélection plus fine pour réduire la charge de calcul.

Dans cet article les données utilisées proviennent de RECS2015, un dataset sur la consommation énergétique des ménages américains en 2015.

## **2 Travaux connexes**

* Étude similaire en Chine révèle que le facteur influençant le plus est le revenu familial avec des corrélations
* Utilisation de l’algorithme d’optimisation des colonies de fourmis et la combinaison de l'algorithme génétique (GA) et de l'ACO (GA-ACO) pour sélectionner les caractéristiques. Les facteurs efficaces sur la consommation d'énergie sont résumés en quatre catégories : météo, heure, économie et perturbations aléatoires.
* Évaluation des performances de 4 méthodes : autocorrelation, mutual information, RReliefF and correlation based. Toutes ont montré leur efficacité.
* Approche en 2 étapes avec l’utilisation de l’algorithme Relief et la cross corrélation pour la sélection de features puis la prédiction avec un réseau neuronal pour la prédiction de la consommation d'énergie.
* Combinaison de filtre-wrapper : élimination des propriétés non pertinentes et redondantes en fonction de la technique de filtrage puis application de l’algorithme Firefly pour éliminer des caractéristiques supplémentaires sans baisser la précision

## **3 Approche proposée**

Sélectionner les caractéristiques les plus pertinentes pour une meilleure compréhension des données, réduction du temps de calcul et amélioration de la performance de précision.

3 catégories de méthodes sélection de features :

* Wrapper : Les méthodes de wrapper mesurent "l'utilité" des caractéristiques en fonction de la performance du classificateur (ex: Naïve Bayes, SVM et K-means)
* Embedded : méthodes utilisées pour optimiser la fonction objectif ou la performance d'un algorithme d'apprentissage (ex: CART et C4.5), la pertinence des features est évaluée pendant l’apprentissage
* Filter : pertinence des features mesurée par des méthodes statistiques (ex: gain d'information, le test du chi-carré, le score de Fisher, le coefficient de corrélation et le seuil de variance)

Deux attributs cibles : "la quantité annuelle de consommation d'énergie en milliers de BTU" (connue sous le nom de TOTALBTU) et "le coût total annuel payé en dollars" (TOTALDOL)

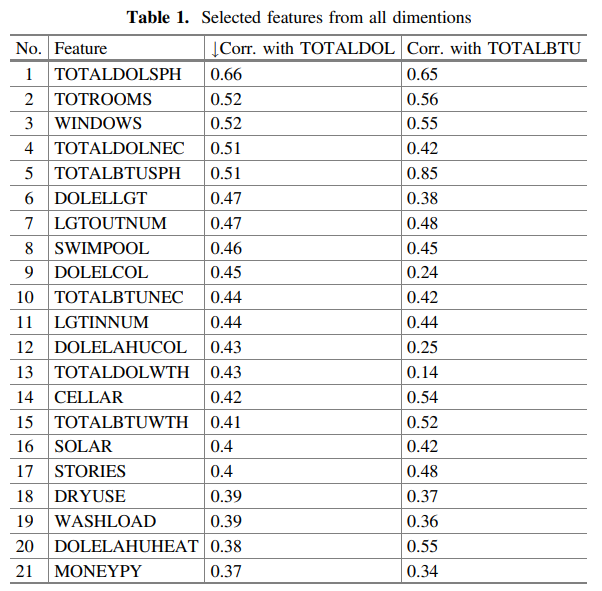
Les méthodes filtres ont été utilisées pour sélectionner les caractéristiques les plus pertinentes en fonction de ces attributs cibles.

**3.1 Dataset**

RECS2015 contient 740 attributs et 5 686 enregistrements de données

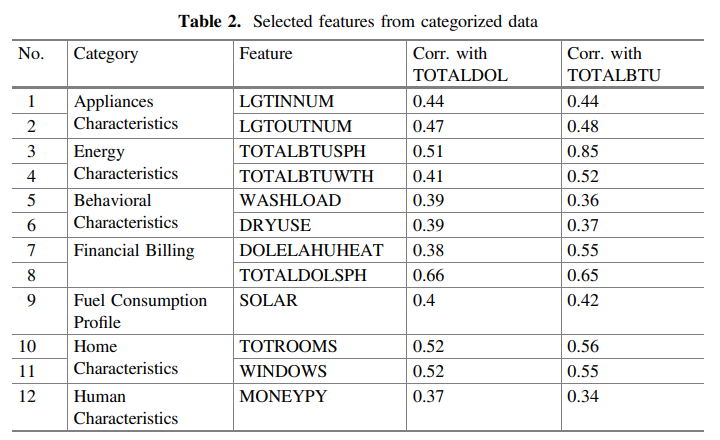
**3.2 Sélection des features**

* Corrélation de Pearson sur toutes les features : calcule la corrélation entre chaque caractéristique et les deux attributs cibles (TOTALBTU et TOTALDOL). Les caractéristiques présentant la corrélation la plus élevée sont sélectionnées



* Corrélation de Pearson sur les données catégorisées : Divisions des features en 8 catégories avant de faire la sélection

Les 12 caractéristiques obtenues par la méthode de corrélation sur les données catégoriques font partie des caractéristiques obtenues lorsque la corrélation est calculée pour toutes les caractéristiques. Cela signifie qu'en catégorisant d'abord les caractéristiques et en sélectionnant ensuite les attributs les plus importants, il est possible d'identifier et de sélectionner les caractéristiques les plus importantes parmi les 740 caractéristiques.

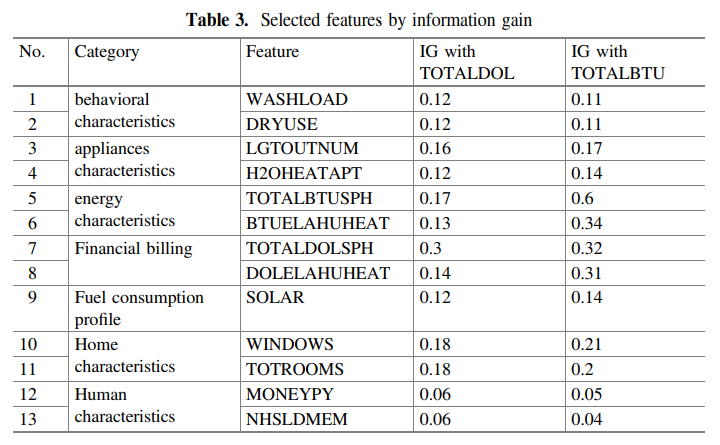


**3.3 Evaluation**

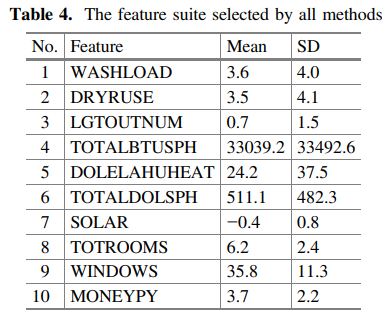
* Gain d’information (IG) : l'un des schémas pour créer un arbre de décision en utilisant la notion d'entropie. Dans cette méthode, à chaque étape, l'attribut qui réduit le plus l'entropie est sélectionné comme nœud de l'arbre (la racine ou le nœud intermédiaire). L'intuition derrière cette méthode est que les attributs qui réduisent davantage l'entropie sont capables de fournir plus d'informations sur les données.

Cette méthode a été effectuée sur chacune des catégories et dans chaque catégorie, les caractéristiques fournissant plus d'informations sur les deux attributs cibles sont sélectionnées comme les plus importantes et efficaces.

Les résultats sont similaires aux deux méthodes précédentes.



L’intersection des résultats des 3 méthodes précédentes :

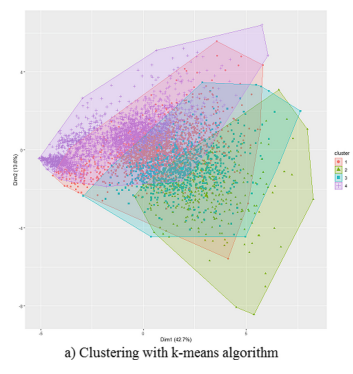


**3.4 Analyse descriptive**

* Clustering basé sur la distance : K-Means

4 clusters

1. Premier cluster : familles qui utilisent fréquemment la machine à laver et le sèche-linge au cours d'une semaine et utilisent un nombre modéré d'ampoules. Les ménages utilisent une quantité d'énergie moyenne, voire inférieure, pour chauffer leur espace de vie, revenu annuel relativement faible (généralement entre 20 000 et 60 000 dollars) -> 27,10% des ménages
2. utilisent une quantité d'énergie plus élevée pour chauffer les espaces de vie et l'eau, revenu annuel élevé (généralement plus de 140 000 dollars). Leurs maisons sont grandes et comportent plus de pièces et de fenêtres -> 5.02% des ménages
3. combinaison des deux premiers clusters -> 15.81% des ménages
4. n'ont généralement pas de machine à laver ni de sèche-linge, et le nombre de lampes utilisées dans leurs maisons est très faible, faible consommation et faible revenu -> 52.05% des ménages

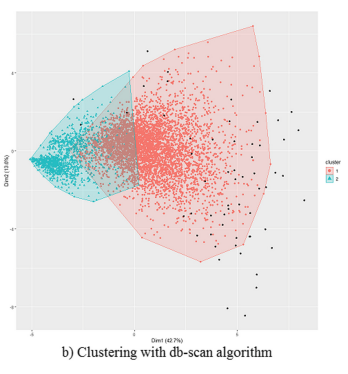


* Clustering basé sur la densité : DB-SCAN

algorithme de clustering basé sur la densité qui est également capable de détecter les valeurs aberrantes - les points de données qui se trouvent dans des régions de faible densité.

2 clusters

1. utilisent une machine à laver et un sèche-linge. Elles utilisent également un nombre assez important de lampes dans leurs maisons. Elles consomment également une quantité d'énergie moyenne à élevée pour chauffer leurs espaces de vie, ce qui entraîne un coût assez élevé à payer pour l'électricité. Le nombre de chambres à coucher et de fenêtres dans les maisons est assez grand. Les ménages de ce cluster ont un revenu annuel moyen à élevé.
2. Les machines à laver et les sèche-linge ne sont pas utilisés. De plus, aucune ampoule n'est utilisée à l'extérieur des maisons. La consommation d'énergie est faible, ce qui entraîne un faible coût à payer. Les maisons de ce cluster ont un petit nombre de chambres à coucher et de fenêtres. Le revenu annuel des familles est très faible.
3. Les ménages dans les données aberrantes comprennent ceux avec un revenu très élevé ou très faible. La quantité d'utilisation des machines à laver et des sèche-linge dans ces deux groupes est très différente, car elles sont utilisées de nombreuses fois au cours de la semaine. La quantité d'énergie consommée dans ces ménages est également différente. En général, les ménages de ce groupe présentent un comportement inhabituel en matière de consommation d'énergie ou leur revenu annuel diffère du reste de la communauté.



En examinant plus en détail les résultats du clustering k-means, on constate que le k-means n'a pas regroupé les données de manière efficace. En fait, seuls les ménages des deux clusters 2 et 4 ont des comportements différents ; les deux autres clusters, c'est-à-dire les clusters 1 et 3, sont une combinaison de ménages pauvres et riches.

En conclusion, les résultats montrent que l'algorithme de clustering DB-SCAN surpasse le k-means dans la partition des données en clusters plus précis, ainsi que dans l'identification de modèles valides de la consommation d'énergie.