
Mise en place d'une méthode de clustering pour la mise en évidence d'un potentiel d'économie d'énergie sur les bâtiments existants.

Martin AMIEL^{1,2}

m.amiel@groupepelletier.fr

¹ LOCIE – Université Savoie Mont Blanc

² Groupe Pelletier

RÉSUMÉ.

Les bâtiments existants et le secteur du bâtiment en général est le premier consommateur d'énergie finale en France et le second émetteur de gaz à effet de serre. Dans ce contexte, les nouvelles constructions se doivent d'être plus respectueuses de l'environnement sur l'ensemble de leur cycle de vie. La nouvelle réglementation environnementale prévue pour 2020 va en ce sens. En ce qui concerne les bâtiments existants, leur impact énergétique et environnemental doit aussi être réduit. Le principal frein à la rénovation est l'investissement. Un des leviers d'actions pour faciliter la rénovation est de générer des économies d'énergies (et donc financières) sans réaliser de travaux et avec un investissement initial limité. Pour cela une méthodologie d'analyse des données de consommations et des usages (issues d'une instrumentation dédiée) est développée pour mettre en évidence ce potentiel d'économie d'énergie en vue d'une rénovation énergétique. Cette analyse, en plus de générer des économies d'énergies permettra une conception de la future rénovation mieux adaptée aux besoins et aux usages présents dans le bâtiment.

ABSTRACT.

Existing buildings and the building sector in general is the largest consumer of final energy in France and the second largest emitter of greenhouse gases. In this context, new construction must be more respectful of the environment over their entire life cycle. The new environmental regulation planned for 2020 goes in this direction. With regard to existing buildings, their energy and environmental impact must also be reduced. The main obstacle to renovation is investment. One of the levers of action is to generate energy savings (and therefore financial) without performing work and with a limited initial investment. For this purpose, a methodology for analyzing consumption data and uses (resulting from dedicated instrumentation) is developed to highlight this potential for energy saving in order to energy renovation. This analysis, in addition to generating energy savings will allow a better design renovation which is more suitable to the needs and uses present in the building.

MOTS-CLÉS : Bâtiments, Diagnostic, Rénovation, Clustering, Economie d'énergie

KEY WORDS: Building, Diagnosis, Renovation, Clustering, Energy saving

1. Introduction

Le secteur du bâtiment, en tant que premier consommateur d'énergie finale en France (45% soit 67.7 Mtep) et second émetteur de CO₂ (23% soit 72 Mt CO₂)¹, doit contribuer au respect des engagements énergétiques et environnementaux pris par la France. Un des leviers que doit utiliser le secteur du bâtiment pour réduire ses impacts est la rénovation énergétique du parc immobilier afin de réduire les consommations liées à son exploitation.

Le terme de rénovation énergétique désigne l'ensemble des travaux réalisés sur un bâtiment visant à diminuer la consommation énergétique du bâtiment et/ou de ses occupants. Dans le cadre de la Loi de Transition Energétique pour la Croissance Verte, un objectif de 500 000 rénovations lourdes par an a été pris. Comme on peut le lire dans le dernier rapport de l'enquête OPEN², en 2014, plus de 288 000 rénovations performantes ont été réalisées dans le secteur privé.

Cependant, pour qu'une rénovation soit efficace, il est nécessaire de bien cerner les besoins des usagers au préalable. Lancer une campagne de travaux sans étudier le comportement du bâtiment et ses usages en amont pourrait annuler les bénéfices des actions de rénovation engagées. Il a été démontré, au travers d'études et d'expérimentations que le seul travail sur les usages des personnes utilisant le bâtiment pouvait générer des économies d'énergies annuelles d'au minimum 10% et dans certains cas jusqu'à 40%, en atteste les résultats de l'expérimentation CUBE 2020³ et [AND 17].

Le travail sur le facteur humain, au travers d'usages responsables a donc un bénéfice non négligeable sur la performance finale du bâtiment. Il est par conséquent important d'analyser et d'exploiter ce potentiel d'économie d'énergie en amont pour que la conception soit en accord avec les usages et le fonctionnement du bâtiment.

Il n'existe, à ce jour, aucun outil capable de mettre en évidence ce potentiel d'économie d'énergie sur un bâtiment en activité. Cela peut s'expliquer de façon assez simple : la mise en évidence de ce potentiel nécessite une connaissance relativement précise à la fois des consommations énergétiques du bâtiment mais aussi de ses usages et cela de façon simultanée. La mise en parallèle de ces deux types de données est indispensable car le potentiel d'économie d'énergie est indexé sur l'usage : on cherche à différencier les périodes où il est normal de consommer celles où non.

L'objectif de ces travaux est donc d'être capable de mettre en évidence des situations de consommation différentes ainsi que les paramètres influents sur ces situations en se basant sur les données de consommations et d'usage d'un bâtiment en activité. Une fois ces regroupements effectués, un potentiel d'économie d'énergie pourra être défini de façon statistique et empirique.

2. Présentation de la méthode

2.1. Pourquoi le clustering

Afin d'être capable de mettre en évidence des situations de consommation identiques, il a été décidé de se rapprocher des méthodes de Data Mining et plus particulièrement du clustering. Le terme de clustering rassemble l'ensemble des méthodes d'analyse de données. Ces méthodes visent à diviser un ensemble de données en différents groupes homogènes. C'est-à-dire que chaque groupe partage des caractéristiques communes. Ces regroupements sont le plus souvent réalisés en introduisant des mesures et classes de distance entre objets.

De façon générale le clustering est utilisé, dans le domaine de l'énergie du bâtiment, pour faire du classement de données de consommation à partir des usages et/ou faire de la prédiction sur les consommations énergétiques du bâtiment [LON 17], [DAV 15], [JUN 17], [CHI 18], [AND 17]. Ici il sera utilisé sur des données d'exploitation d'un bâtiment.

Parmi les différentes méthodes de clustering, on retrouve :

- Méthodes des centroïdes : pour un nombre de cluster K fixe, lors de l'initialisation, K positions moyennes sont définies. On affecte ensuite chaque observation au cluster le plus proche (minimisation de la distance intra-cluster). Une fois les différents regroupements réalisés, on recalcule les positions moyennes des différents clusters et réitère l'affectation jusqu'à la convergence. Pour un nombre de cluster K fixé, il y a

¹ Chiffres de l'énergie 2016, Service de l'observation statistique du Ministère de l'Ecologie

² https://www.ademe.fr/sites/default/files/assets/documents/open_2015_8679.pdf

³ <https://cube2020.org/resultats>

un nombre de fini de partitions possible ce qui assure la convergence du résultat. Il est cependant possible que le temps de calcul devienne important.

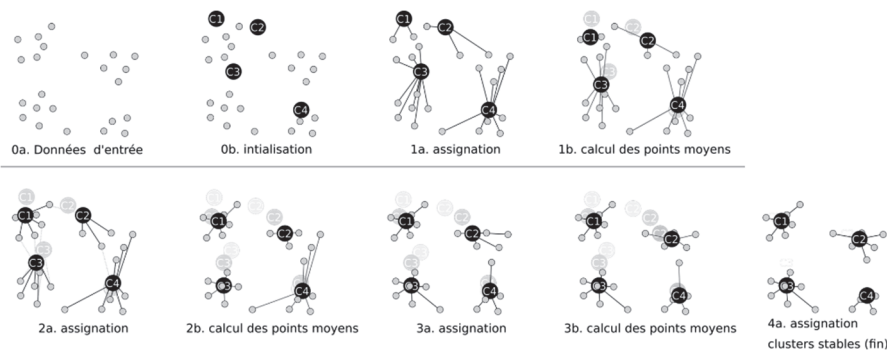


Figure 1 : Illustration du déroulement de l'algorithme des k-moyennes (Source : Wikipédia)

- Méthode de regroupement hiérarchique : pour un ensemble de n individus, initialement chaque individu forme un groupe. A chaque itération, les groupes ayant l'indice de dissimilarité le plus faible sont regroupés deux à deux. Lors de la dernière itération il n'existe plus qu'un seul groupe. L'indice de dissimilarité est basé sur les distances entre individus.

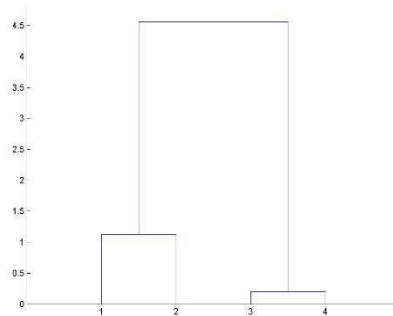


Figure 2 : Illustration du déroulement de l'algorithme de regroupement hiérarchique (Source : Wikipédia)

- Algorithme de maximisation de l'espérance : L'algorithme d'espérance-maximisation comporte :
 - o Une étape d'évaluation de l'espérance (E), où l'on calcule l'espérance de la vraisemblance en tenant compte des dernières variables observées,
 - o Une étape de maximisation (M), où l'on estime le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape E.

On utilise ensuite les paramètres trouvés en M comme point de départ d'une nouvelle phase d'évaluation de l'espérance, et l'on itère ainsi.

- Algorithme de densité : ce sont des algorithmes réalisant des regroupements en se basant sur la distance entre individus et un nombre minimal de point par regroupement. A partir d'un point donné, on parcourt son voisinage pour vérifier ces hypothèses et réaliser des regroupements.

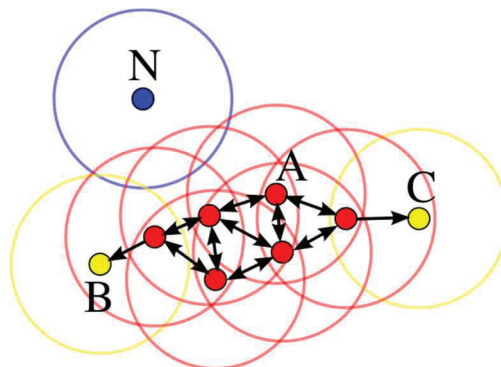


Figure 3 : Illustration du déroulement de l'algorithme de densité (Source : Wikipédia)

Pour les besoins de cette étude, il a été choisi d'utiliser l'algorithme des k-moyennes (k-means). Cette méthode de clustering a déjà été utilisée dans un grand nombre d'études sur la performance énergétique des bâtiments [MAH 17], [JIA 15], [YUN 16]. Reconnue et approuvée par la communauté scientifique, cet algorithme, permet une grande modularité et une implémentation facile et rapide. On ne sait pas combien de clusters doivent être réalisés ni même le nombre de paramètres à prendre en compte pour réaliser les regroupements. L'apprentissage non-supervisé est donc parfaitement adapté aux besoins de cette étude.

2.2. Présentation de la méthode de clustering

L'objectif de la méthode développée est de mettre en évidence des situations de consommation différentes en réalisant des regroupements de données de consommation et d'usage du bâtiment. Pour tester et valider la méthode, des simulations thermiques dynamiques vont être réalisées afin de disposer de jeux de données complets et fiables. Différents scénarios d'usages ont été modélisés : un scénario « type » censé représenter le bâtiment lors d'un fonctionnement « normal » et différents scénarios représentants des dérives (hausse du thermostat, ouvertures de fenêtre, modification de la régulation...). A l'issue de ces différentes modélisations, plusieurs types de données sont mises à disposition :

- Consommations énergétiques : chauffage, éclairage, eau chaude sanitaire, électricité spécifique, ventilation
- Données d'usage : température intérieure, hygrométrie intérieure, occupation, ouverture de fenêtre
- L'environnement extérieur : ensoleillement, température, hygrométrie, vitesse du vent

L'ensemble de ces données sont fournies au pas de temps journalier.

Pour réaliser les différents regroupements, aucune connaissance à priori n'est affectée sur les paramètres ni sur le nombre de clusters à réaliser. C'est-à-dire qu'on ne sait pas quels sont les paramètres influents ni même le nombre de clusters optimal nécessaire pour représenter au mieux les modalités de consommations. En ce sens, toutes les combinaisons de paramètres, suivant un nombre de clusters variable, seront évaluées.

Soit P le nombre de paramètres, N le nombre d'observations et K le nombre de clusters, l'ensemble des combinaisons testées respecteront les deux conditions suivantes :

$$K > 1 \quad [1]$$

$$K < P \quad [2]$$

Le choix de ces deux conditions s'explique simplement :

- Un seul cluster ne permet pas de représenter différentes situations de consommations, plus généralement un nombre de clusters trop faible rendrait ces clusters trop généralistes contenant beaucoup de données
- Si l'on a un nombre de clusters trop important, on risque d'avoir un partitionnement trop fragmenté des données et on passerait à côté de certains comportements

Afin de ne pas biaiser le clustering, pour chaque paramètre, les observations seront centrées réduites. L'algorithme des k-means, réalisant des regroupements à partir des distances entre observations ce pré-requis est indispensable. Les différents paramètres utilisés dans cette étude n'ont pas les mêmes ordres de grandeur. Si les variables ne sont pas centrées réduites, les variables avec la valeur numérique la plus grande auraient un impact plus important que les autres.

Une fois l'ensemble des différentes combinaisons testées, il est nécessaire de choisir celle qui représente le mieux le comportement du bâtiment. Pour rappel, la méthode des k-means vise à minimiser (pour un nombre de clusters K fixé) la distance entre le centroïde d'un cluster et les observations incluses dans ce même cluster, la somme de ces valeurs est équivalente à la variance du cluster. Pour mesurer le bénéfice du regroupement en K cluster, on calcule le ratio entre la somme des variances de chaque cluster et la variance de l'échantillon. Ce coefficient est l'équivalent du R^2 pour les régressions linéaires. Dans la suite du document il sera appelé indicateur de performance (toujours inférieur à 1 et exprimé en pourcent). Plus cet indicateur est élevé plus les clusters formés sont censés représenter fidèlement le comportement du bâtiment.

Afin d'aider à la décision, un second indicateur est calculé : l'écart-type des centroïdes des clusters créés pour chaque combinaison. Il permet de représenter la distance entre les différents centroïdes. Plus cette valeur est élevée plus les clusters composés représentent des situations de consommation différentes (maximisation de la distance inter-cluster). Cet écart-type est calculé sur la consommation globale du bâtiment (tous usages confondus) qui est le paramètre de sélection.

Une fois ces deux indicateurs calculés pour toutes les combinaisons, les résultats sont classés par valeur de K et suivant leur indice de performance (valeur décroissante). Lorsque le nombre de paramètres P devient grand, le nombre de combinaisons peut vite devenir important. Le nombre total de combinaisons pour P paramètres est de :

$$\sum_{j=2}^{P-1} \sum_{i=3}^P \binom{i}{j} \quad [3]$$

Pour plus de lisibilité, seules les 10 premières combinaisons pour K fixé sont présentées : cela permet de visualiser rapidement les meilleures combinaisons. A partir des différents résultats, l'opérateur peut choisir le regroupement qui lui semble le mieux adapté à la situation. A l'issue de cette sélection, la distribution de chaque cluster suivant les jours de la semaine est affichée, et il est laissé à l'opérateur le choix de supprimer certaines observations s'il juge qu'elles ne sont pas représentatives (c.à.d si un jour est significativement moins représenté au sein du cluster que les autres).

A titre indicatif, il est aussi possible de visualiser, parmi les 10 meilleures combinaisons, pour chaque valeur de K, les paramètres les plus souvent appelés.

L'ensemble de la méthodologie a été développée sous R à l'aide de la fonction « kmeans » du package *stats*.

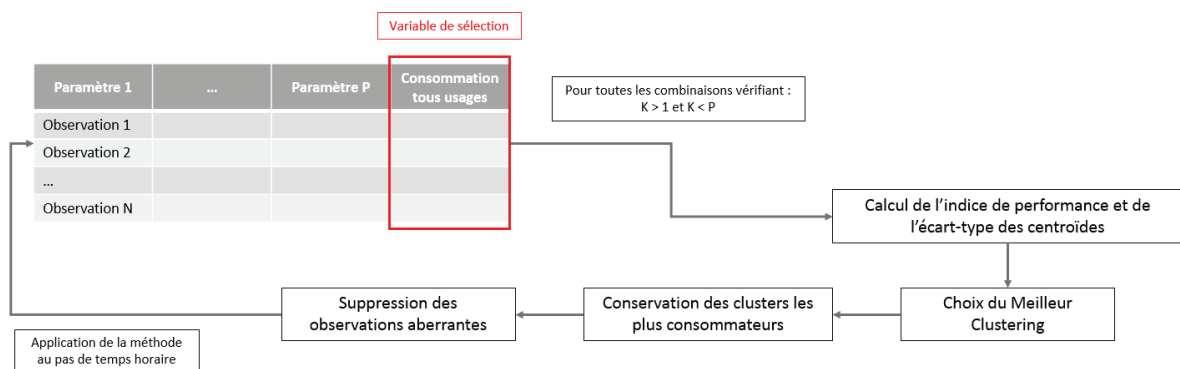


Figure 4 : Illustration de la méthodologie de clustering

Cette démarche s'effectue dans un premier temps au pas de temps journalier. A l'issue de ce clustering, nous avons donc des regroupements de journées où les profils de consommation sont identiques. Cependant, le niveau de détail à cette échelle n'est pas jugé suffisant pour mettre en évidence un potentiel d'économie d'énergie.

Parmi les clusters mis en évidence, seuls ceux représentant plus de 1/K des consommations sont retenus. Dans les autres cas, il est considéré que la part des consommations est trop faible pour générer des économies d'énergies. Pour les clusters retenus lors de cette première étape, la méthode précédemment décrite leur est appliquée mais cette fois au pas de temps horaire. Il a été décidé de réappliquer la méthode dans sa totalité et non seulement reprendre le même nombre de clusters K et les paramètres influents car les modalités de consommation peuvent être différentes lorsque l'on passe à un niveau de détail plus élevé.

2.3. Application de la méthode de Clustering

La méthode de clustering présentée précédemment a été appliquée sur un bâtiment instrumenté en exploitation. L'application de la méthode a permis de connaître les variables les plus influentes sur les consommations.

Le bâtiment étudié est une Médiathèque (déjà rénovée) située à Montmélian en Savoie. Son système de chauffage est une chaudière gaz réalisant un appoint à la batterie chaude de la centrale de traitement d'air assurant le renouvellement d'air intérieur. L'éclairage est assuré par des panneaux LED sur détection de présence.

Parmi les données mesurées, nous avons :

- La consommation de chauffage [kWh]
- La température intérieure [°C]
- La température extérieure [°C]
- La consommation de la centrale de traitement d'air [kWh]
- Les consommations d'éclairages [kWh]
- Les consommations des prises (usages spécifique de l'électricité) [kWh]

	Indice de performance	Ecart-type	Chauffage	T_int	T_ext	CTA	Eclairage	Prises	Occupation
K=2	58,38	0,27				X	X	X	
	54,76	0,95	X	X	X				
	53,83	1,25	X		X	X			
	51,36	0,85		X	X	X			
	49,13	1,34	X		X		X		
	48,70	0,93	X	X	X	X			
	47,75	0,20				X	X		X
	47,35	1,37	X		X			X	
	47,31	1,40	X		X				X
	47,16	0,20				X	X	X	X
K=3	64,34	1,23	X	X	X	X			
	61,28	1,22	X		X		X		X
	60,33	0,24				X	X	X	X
	60,29	1,26	X	X	X		X		
	59,64	1,31	X				X	X	X
	59,24	1,23	X		X		X	X	
	59,05	1,15	X		X	X	X		
	58,76	1,25	X	X	X			X	
	58,75	1,25	X	X	X				X
	58,06	1,21	X		X	X			X
K=4	64,37	1,12	X	X	X		X		X
	64,37	1,11	X		X		X	X	X
	63,66	1,05	X	X	X	X	X		
	63,06	1,07	X		X	X	X		X
	62,67	1,14	X	X	X		X	X	
	62,61	1,13	X			X	X	X	X
	62,47	1,08	X		X	X	X	X	
	61,54	0,69		X	X		X	X	X
	61,48	1,11	X	X	X	X		X	
	61,23	1,08	X	X	X	X			X
K=5	65,53	1,05	X		X	X	X	X	X
	65,11	1,05	X	X	X		X	X	X
	64,80	1,05	X	X	X	X	X	X	
	64,16	1,03	X	X	X	X	X		X
	62,93	0,71		X	X	X	X	X	X
	62,44	1,06	X	X	X	X		X	X
	61,84	1,04	X	X	X	X	X	X	X
60,59	0,93	X	X		X	X	X	X	

Figure 5 : Occurrence des paramètres suivant le nombre de cluster

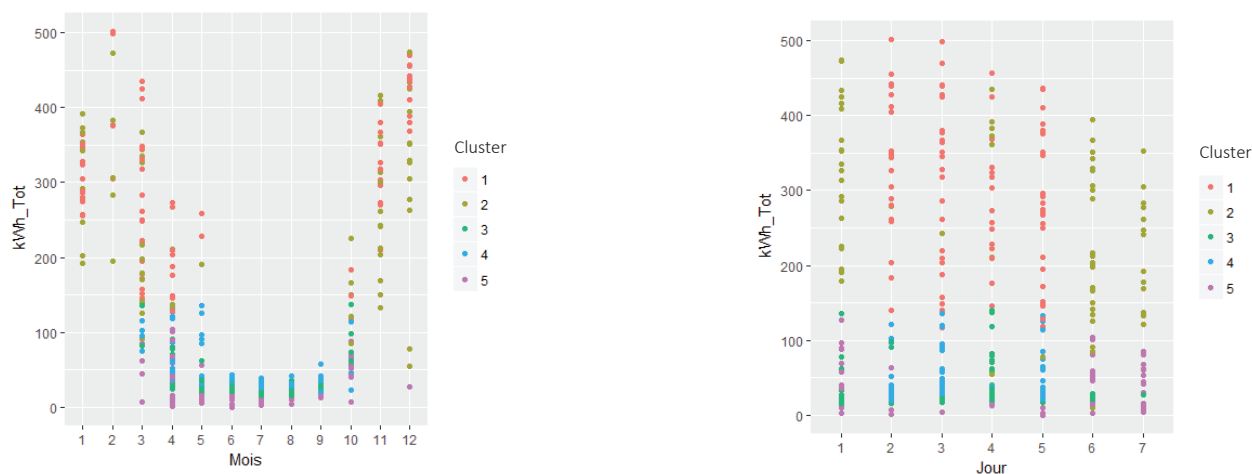


Figure 6 : Répartition des clusters suivant le Mois et les Jours de la semaine

Les données utilisées ici sont celle de la saison 2017/2018. Le regroupement choisi dans ce cas, utilise 5 clusters. Parmi les paramètres retenus, on retrouve la consommation de chauffage, la température extérieure, les consommations de la centrale de traitement d'air, de l'éclairage et des prises. Ce regroupement permet de mettre en évidence les phénomènes de saisonnalité mais aussi d'usage au sein de la semaine (Figure 6).

De manière plus générale, les paramètres les plus représentés dans ce cas précis sont les consommations de chauffage et la température extérieure (Figure 5). Ces résultats sont cohérents car le bâtiment vient d'être rénové et possède une régulation adaptée aux usages du bâtiment. Dans le cas d'un bâtiment existant et ancien on peut s'attendre à ce que les résultats soient bien différents et laissent une place plus importante aux usages électrique et la température intérieure.

3. Conclusion

Cette méthodologie est encore en cours de validation mais montre déjà des résultats prometteurs. L'utilisation des simulations thermiques dynamiques en cours de réalisation pour la validation de la méthode de clustering viendront confirmer ces résultats et serviront par la même occasion au développement de la méthode de mise en évidence du potentiel d'économie d'énergie.

La segmentation de la base de données pour l'étude de clustering est aussi une amélioration possible de la méthode. Cela aura pour effet ne pas s'affranchir des phénomènes de mi-saison et de travailler sur des périodes homogènes du point de vue de l'environnement extérieur réduisant ainsi l'impact de ce dernier et favorisant l'impact des usages au sein de la méthode de clustering.

Cette méthodologie est basée sur une étude statistique des clusters les plus consommateurs avec pour objectif de différencier les périodes où il est normal de beaucoup consommer et celles où non pour finalement en tirer un potentiel d'économie d'énergie.

Plus les usages sont faibles et les consommations élevées plus le potentiel est important et inversement. Ce principe est présenté dans la Figure 7 (ci-dessous).

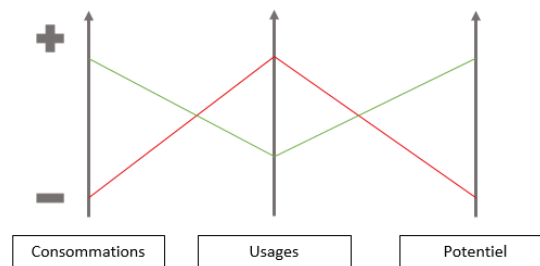


Figure 7 : Schéma de principe de la méthode de mise en évidence du potentiel d'économie d'énergie

4. Bibliographie

- [LON 17] LONGQUAN D., YONJUN S., ZEJUN C., JIAYU C., « Modeling energy consumption in residential buildings: A bottom-up analysis based on occupant behavior pattern clustering and stochastic simulation », *Energy and Buildings 147* (2017), p. 47-66.
- [DAV 15] DAVID H., « Comparaison of integrated clustering methods for accurate and stable prediction of building energy consumption data », *Applied Energy 160* (2015), p. 153-163.
- [JUN 17] JUNJING Y., CHAO N., CHIRAG D., FAN Z., DAVID C., SIEW E-L., CHANDRA S., KWOK W-T., « k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement », *Energy and Buildings 146* (2017), p. 27-37.
- [CHI 18] CHIRAG D., SIEW E-L., « Determining key variables influencing energy consumption in office buildings through cluster analysis of pre and post-retrofit building data », *Energy and Buildings 159* (2018), p. 228-245.
- [AND 17] ANDRII Z., AXEL S., « Building energy data analysis by clustering measured daily profiles », *CISBAT 2017 International Conference – Future Buildings and Districts – Energy Efficiency from Nano to Urban Scale*, Lausanne, 6-8 September 2017, *Energy Procedia 122* (2017) p. 583-588.
- [MAH 17] MAHER A., FREDIK W., « Smart meter data clustering using consumption indicators: responsibility factor and consumption variability », *ICAE2017 - 9th International Conference on Applied Energy*, Cardiff, 21-24 August 2017, *Energy Procedia 142* (2017), p. 2236-2242.
- [JIA 15] JIANPING W., JIANING Z., « Evaluation on building end-user energy consumption using clustering algorithm », *9th International Symposium on Heating, Ventilation and Air Conditioning (ISHVAC) and the 3rd International Conference on Building Energy and Environment (COBEE)*, *Procedia Engineering 121* (2015), p. 1144-1149.
- [YUN 16] YUNSECK L., « Application de techniques de clustering pour la segmentation de parcs de bâtiments à rénover », *Journée des Doctorants – CSTB*, Paris, 4 Octobre 2016.
- [ADN 17] ADNANE K., NATHALIE L., KÉVIN M., « What do people ‘Learn by looking’ at direct feedback on their energy consumption? Results of a field study in southern France », *Energy Policy 108* (2017), p. 593-805.