Méthode statistique de classification automatique des mesures de consommations de bâtiments d'enseignement en eau, gaz et électricité pour la mise en évidence d'anomalies de fonctionnement

Mostafa Akil, Didier Defer, Pierre Tittelein, Frédéric Suard

Méthode statistique de classification automatique des mesures de consommations de bâtiments d'enseignement en eau, gaz et électricité pour la mise en évidence d'anomalies de fonctionnement

Mostafa AKIL*[1], Didier DEFER[1], Pierre TITTELEIN[1], Frédéric SUARD[2]

[1] Laboratoire de Génie Civil et géo-Environnement LGCgE, Université d'Artois, Pôle Béthune, Faculté des Sciences Appliquées, Technoparc Futura, 62400 Béthune, France

[2] LIANES - Laboratoire Intelligence Artificielle pour les Energies NouvelleS, Département des Technologies Solaires - CEA/Liten, Département Métrologie, Instrumentation et Information – CEA/List Lynx 3, 50 avenue du Lac Léman -| F-73375 Le Bourget-du-Lac

*** mostafa_akil@ens.univ-artois.fr**

RESUME. Les gestionnaires de parc équipent de plus en plus leurs bâtiments avec de nombreux capteurs. L'analyse des données issues de ces mesures reste pourtant compliquée et se limite souvent à la définition de seuils d'alerte. Cet article vise à développer une méthodologie d'analyse basée sur l'élaboration d'indicateurs statistiques. Leur suivi pourra permettre de détecter des évolutions de comportement ou de performance des systèmes. La mise en évidence automatisée d'anomalies de fonctionnement pourra déclencher des alertes. Notre première approche se concentre sur le traitement de données issues de collèges gérés par le département de Pas de Calais. 117 bâtiments de collège sont instrumentés avec différents capteurs (eau, électricité, gaz, température intérieure et extérieure, …) et fournissent des données depuis 2015. Plusieurs cycles de saisons sont exploitables.

Des méthodes de Data Mining, y compris l'approche de Clustering, ont été utilisées pour extraire des informations à partir des mesures en 2015 et 2016. Les résultats de la classification des données ont permis de mettre en évidence des jours de fonctionnement rares en exploitant les résultats de la classification des données.

MOTS-CLÉS : Data Mining, K-Means, Arbre de décision, données énergétiques, maintenance de bâtiment.

ABSTRACT. Fleet managers are increasingly equipping their buildings with many sensors. The analysis of the data resulting from these measurements remains complicated and is often limited to the definition of alert thresholds. This article aims to develop an analysis methodology based on the development of statistical indicators. Their monitoring will make it possible to detect changes in the behavior or performance of the systems. The automated highlighting of malfunctions and changes may trigger alerts. Our first approach focuses on data processing from colleges managed by the department of Pas de Calais. 117 college buildings are instrumented with different sensors (water, electricity, gas, indoor and outdoor temperature…) and provide data since 2015. Several seasons cycles are exploitable

Data Mining methods, including the Clustering approach, were used to extract information from the measurements in two colleges in 2015 and 2016. The results of the data classification were able to detect rare days of operation. exploiting the results of the classification of the data.

KEYWORDS : Data Mining, K-Means, decision tree.

1. INTRODUCTION

Le département du Pas-de-Calais gère 117 bâtiments de collège. Un plan d'instrumentation a été mis en œuvre depuis plusieurs années pour aboutir au suivi en quasi temps réel des consommations (eau, électricité, gaz) et des températures (intérieure, extérieure). La simple définition de seuils d'alerte sur les consommation d'eau a déjà permis de faire des économies très conséquentes en détectant rapidement

les fuites. L'analyse des autres mesures est plus lourde et complexe, notamment par la dépendance aux usages et à l'environnement. Cet objectif relève des techniques de traitement statistique des données.

Ces techniques ont déjà été fréquemment utilisées pour soutenir et améliorer les aspects fondamentaux de la gestion de l'efficacité énergétique. Citons par exemple Yu et al. (Yu et al. 2010) qui ont proposé l'utilisation d'arbres de décision pour développer des modèles prédictifs de la demande d'énergie de constructions puisqu'ils sont plus facilement interprétés que d'autres techniques de classification. Xaio et Fan (Xiao et Fan 2014) ont utilisé le Clustering pour identifier les modèles de consommation d'énergie quotidienne. Morbitzer et al. (Morbitzer, Strachan, et Simpson 2004) ont également appliqué des algorithmes de Clustering pour traiter les données de surveillance des bâtiments et découvrir des facteurs non évidents de surconsommation d'énergie dans les infrastructures du bâtiment. Capozzoli et al. (Capozzoli, Lauro, et Khan 2015) décrivent une approche simplifiée pour détecter automatiquement les défauts dans les équipements énergétiques du bâtiment. Chicco et al. (Chicco et al. 2004), regroupent les clients en classes, en fonction de leur comportement de consommation électrique.

L'idée est donc ici de développer une méthodologie d'analyse basée sur l'élaboration d'indicateurs statistiques. Leur suivi pourra permettre de détecter des évolutions de comportement ou de performance des systèmes. La première étape de la démarche consiste à se focaliser sur chaque collège indépendamment des autres en vue de combiner les décisions dans la deuxième étape. Elle se focalise sur les mesures de consommation en eau, gaz et électricité. Le but est d'exploiter ces données afin d'élaborer des indicateurs calculés automatiquement, pour mettre en évidence des comportements inhabituels et détecter les anomalies de fonctionnement.

2. METHODE

Notre démarche de travail est d'extraire des informations des consommations quotidiennes des collèges fournis par les systèmes de télé-relève sur la période 2015 et 2016 (eau, électricité et gaz) pour tenter de mettre en évidences d'éventuelles anomalies. Ce travail est initié sur un des collèges notés par la suite « collège B ». Le graphe (Fig 1) affiche le calendrier de 2015 en représentant verticalement les 7 jours de chaque semaine de l'année. Certaines périodes sont inexploitables à cause de problèmes de télérelève (le 17/03/2015 et la période du 26/10/2015 au 31/10/2015). Elles ne concernent cependant que moins de 2% sur l'ensemble des données, la qualité des données ne nécessite donc pas de traitement correctif approfondi.

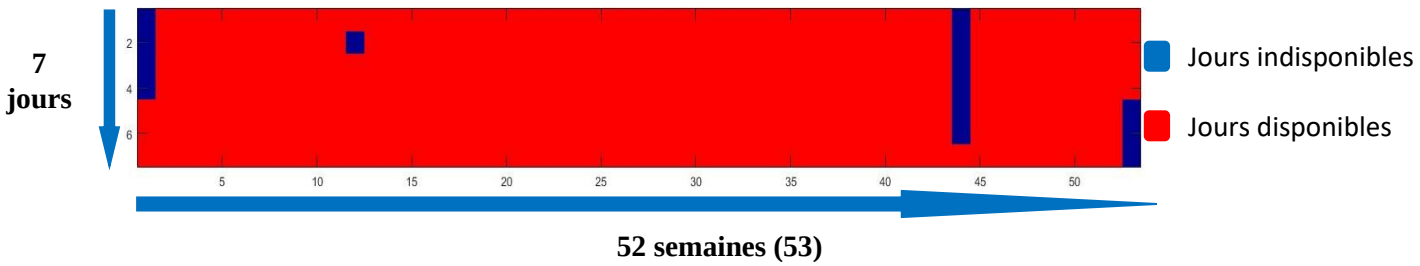

Figure 1 : Calendrier des jours disponibles et indisponibles pour le collège B en 2015

Les consommations sont généralement étudiées quotidiennement. Pourtant dans le mode de fonctionnement des collèges, on peut distinguer des jours complets d'activité (lundis, mardis, jeudis, vendredis), un fonctionnement par demi-journées les mercredis et des jours complets d'inactivité les week-ends. Ceci a conduit à opter pour un découpage de chaque jour en 3 périodes :

- La nuit : de 19h la veille à 5h du matin (pour ne pas inclure la relance du chauffage)
- Le matin : de 5h à 12h
- L'après-midi : de 12h à 19h

Cette répartition a été établie après concertation avec le gestionnaire et la prise en compte de la programmation du scénario de consigne de chauffage. Les données mesurées au pas d'acquisition de 10 minutes sont cumulées en niveaux de consommation dans chaque période aboutissant ainsi à 9 données par jour : 3 pour la consommation d'eau, 3 pour l'électricité et 3 pour le gaz. On doit donc traiter 9 séries de données sur toute l'année.

L'objectif du travail est de proposer un modèle automatisable qui puisse être appliquée à l'ensemble des 117 collèges. Notre première étape proposée, est de trouver une méthode capable de regrouper les jours en familles selon leurs niveaux de consommation. Mais le problème c'est qu'à priori, aucune information n'existe dans les données sur les consommations des 3 pôles dans chaque étape de la journée. Pour cela une approche non supervisée qui permet d'exploiter les données sans information a priori est privilégiée. La méthode de clustering K-Means qui a été retenue vise à regrouper dans des familles (classes), les individus (séries de mesures sur une période donnée dans notre cas) qui se « ressemblent ». Dans ce cas, elle va définir des classes de consommation.

Soit k le nombre de classes dans lesquelles, on souhaite répartir les données. La procédure mise en œuvre dans la méthode K-Means (Usman, Ahmad, et Ahmad 2013)suit la séquence suivante :

1. Etape initiale : Dans l'ensemble des éléments à classer, un tirage aléatoire permet de choisir au hasard k éléments constituant les centres initiaux des k classes.

2. Affectation dans les classes : Chaque individu ou élément de l'ensemble est affecté à la classe dont il est le plus proche du centre. Différentes mesures permettent de quantifier cette proximité. La distance euclidienne retenue dans ce travail exprime la distance entre l'élément $x_i^{(j)}$ et la classe c_j par $\left\| x_i^{(j)} - c_j \right\|^2$.

La fonction objectif J que l'on doit minimiser vise à minimiser la dispersion de chaque classe et s'écrit comme la somme des distances de chaque élément au centre de classe auquel il est rattaché

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} \left\| x_i^{(j)} - c_j \right\|^2 \quad \textbf{(1)}$$

3. Dans chaque classe qui vient d'être constituée, on détermine le nouveau barycentre qui devient le nouveau centre de la classe. Deux cas sont possibles :

    - Les k nouveaux centres de classes sont inchangés et alors la classification est terminée

    - Le nouveau groupe de centres de classe est modifié et alors on réitère la phase 2 d'affectation avec le nouveau groupe.

A l'issue de ce processus itératif, les k classes sont déterminés.

Le nombre k de classes doit être fixé. Dans le cadre d'une procédure automatisée de traitement, c'est-à-dire n'impliquant pas de paramétrage par un utilisateur non expert, le critère Davies-Bouldin (DB) a été utilisé pour retenir le nombre optimal de classes. Il se calcule de la façon suivante :

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max_{j \neq i}\left\{\frac{I(C_i)+I(C_j)}{I(C_i,C_j)}\right\} \quad \textbf{(2)}$$

Pour chaque classe i de la partition, on cherche la classe j qui maximise l' « indice de similarité » décrit comme suit :

$$R_{ij} = \frac{I(C_i)+I(C_j)}{I(C_i,C_j)} \quad \textbf{(3)}$$

$I(C_i)$, représente la moyenne des distances entre les individus appartenant à la classe C_i et son centre alors que $I(C_i, C_j)$ représente la distance entre les centres des deux classes C_i et C_j. La meilleure partition est donc celle qui minimise la moyenne de la valeur calculée pour chaque classe. En d'autres termes, la meilleure partition est celle qui minimise la similarité entre les classes (Davies et Bouldin 1979).

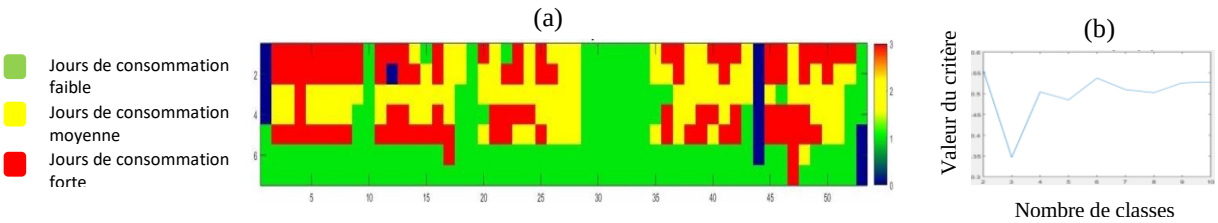

Figure 2 : (a) classification automatique avec K-Means de consommation d'eau le matin de B en 2015, (b) Nombre de classe optimal pour la classification automatique avec DB.

La figure 2 montre (a) le résultat d'une classification par K-Means des consommations d'eau pour la période du matin. Le critère de DB (b) est ici minimisé pour une répartition en 3 classes. Chaque classe « famille » comporte les jours qui se ressemblent par leurs niveaux de consommation d'eau sans définir a priori de niveau de seuil pour passer d'une classe à l'autre. Une couleur a été associée à chaque classe : vert pour la famille des jours de consommation faible, jaune pour une consommation moyenne, et rouge pour une consommation classée forte. On obtient une représentation équivalente pour chacune des 9 séries de données.

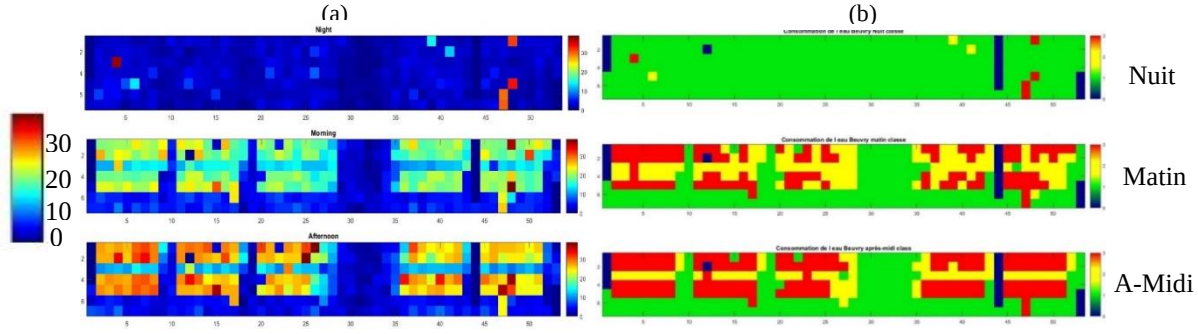

Figure 3 : (a) Calendriers de valeurs brutes de consommation d'eau, (b) leur regroupement avec la classification automatique K-Means de B en 2015.

Sur la figure 3 (a), les calendriers représentent les consommations d'eau codées sur une échelle de couleur allant du bleu pour les plus faibles consommations au rouge pour les plus fortes. Sur la partie de droite (b), on observe le résultat de la classification réalisée par K-Means. Les jours du calendrier de nuit en (a) qui se ressemblent par leur couleur bleue, sont regroupé dans une famille ensemble par K-Means et ils sont représentés par une couleur verte en (b). Alors, en général, la classe des jours de consommations faibles est représentée en vert, la classe moyenne en jaune et la classe forte en rouge. Visuellement, on peut entamer une analyse qui permet de distinguer les jours d'activité, les week-ends, les périodes de vacances, les mercredis. On constate logiquement que la consommation quotidienne d'eau est très liée à la présence des étudiants dans le collège. La consommation d'eau est toujours supérieure à zéro. Pendant les périodes de vacances, cette consommation diminue pour atteindre son minimum.

On peut aussi observer des pics de consommation d'eau durant certaines nuits dans le collège. La consommation d'eau l'après-midi est généralement supérieure à celle du matin. On constate que la présence le mercredi est inférieure à celle des autres journées de la semaine même le matin. On visualise notamment les week-ends où l'activité est réduite (2 rangées du bas de chaque calendrier).

Si on croise les classifications des 3 pôles de consommation, on peut remarquer des combinaisons rares qui peuvent conduire à entamer une analyse plus fine et experte des consommations correspondantes.

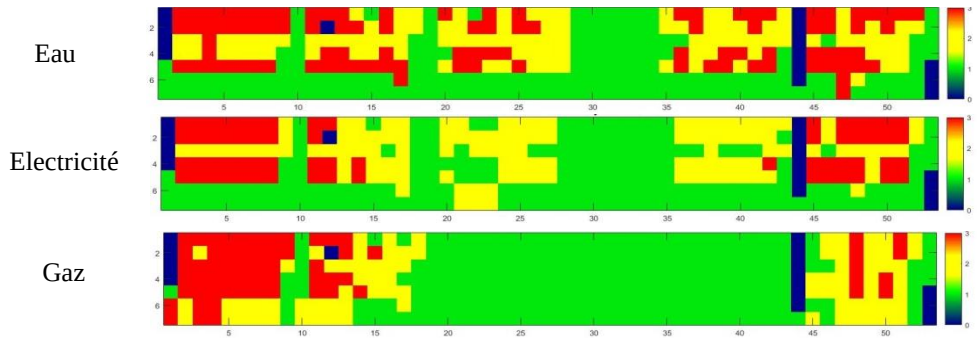

Figure 4 : classification automatique avec K-Means des consommations d'eau, d'électricité et du gaz le matin en 2015

Sur la figure 4 on peut observer par exemple, en bas à gauche, le week-end de la première semaine de l'année 2015. Dans ce cas, on a une forte consommation du gaz combinée avec une faible consommation d'eau et d'électricité. Il s'agit d'une situation qui ne se retrouve que rarement les autres week-ends. Ces jours particuliers nécessiteraient une analyse plus poussés. Cette approche visuelle est intéressante mais elle est difficilement généralisable à l'ensemble des bâtiments instrumentés. Pour automatiser l'analyse, le clustering est complété par la construction d'un arbre de décision qui va combiner les résultats obtenus sur les classifications des 3 pôles de consommation, dans le but de trouver des proportions pour chaque situation de fonctionnement des jours et définir après celles de statuts rares et celles de statuts fréquents.

Un arbre de décision est un classificateur non paramétrique. Le processus de construction de l'arbre de décision est présenté dans (Otukei et Blaschke 2010). La structure de base de l'arbre de décision

consiste en un nœud racine, un certain nombre de nœuds internes et enfin un ensemble de nœuds terminaux. Les données sont réparties récursivement dans l'arbre de décision selon le cadre de classification défini. A chaque nœud, une règle de décision est requise et ceci peut être implémenté en utilisant un test de division souvent de la forme

$$F(\mathbf{x}) = \begin{cases} \mathbf{x_j} > \boldsymbol{c} & \text{pour les arbres de décision univariés} \\ \sum_{j=1}^{n} a_j \mathbf{x_j} \leq \boldsymbol{c} & \text{pour les arbres de décision multivariés} \end{cases} \quad \textbf{(4)}$$

Où $\mathbf{x_j}$ représente les vecteurs de mesure sur les n entités sélectionnées et a_j est un vecteur de coefficients de discrimination linéaire alors que $\boldsymbol{c}$ est le seuil de décision (Otukei et Blaschke 2010).

3. RESULTATS ET DISCUSSION

La figure 5 représente le résultat du traitement des consommations d'électricité du matin obtenu automatiquement pour le collège en inférant des règles de décision sur les autres variables descriptives : eau et gaz. On aboutit à un arbre de décision qui permet de repérer les combinaisons de classification observables de façon rare ou fréquente.

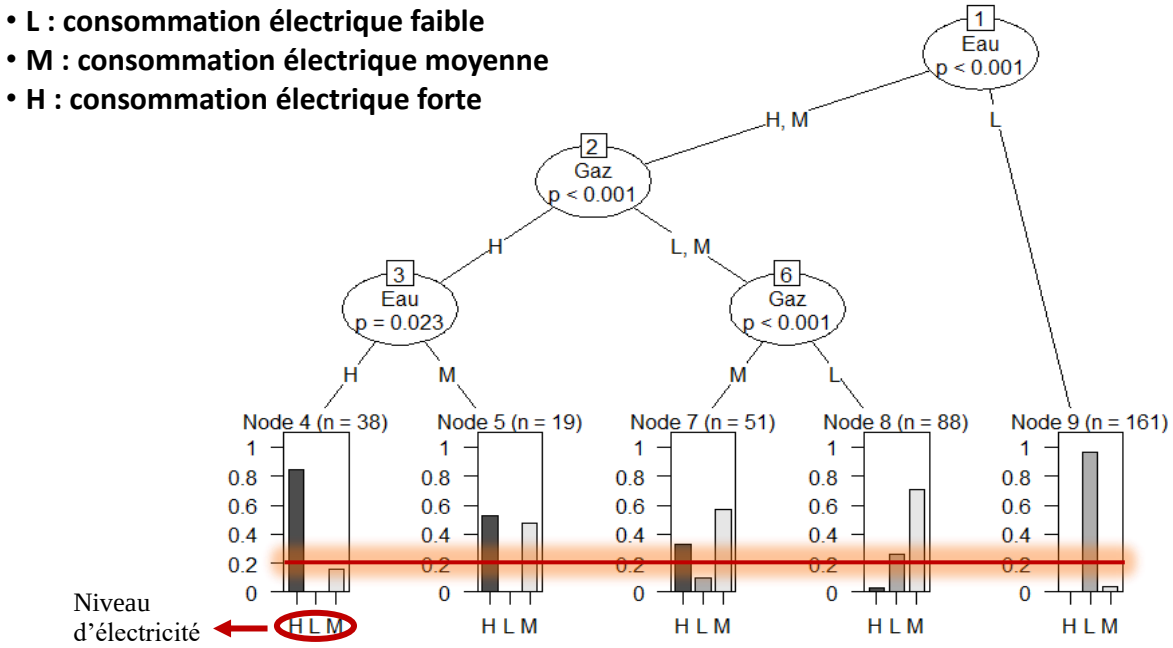

Figure 5 : arbre de décision pour le matin de l'année 2015

D'après la figure 5, certaines combinaisons de consommation apparaissent dans de faibles proportions. Par exemple en prenant la première branche tout à droite de l'arbre, on a observé 161 matinées de 2015 dans le collège B présentant une faible consommation d'eau. Parmi ces individus statistiques, presque 98% sont associés à un faible niveau de consommation en électricité et ce, quel que soit le niveau de consommation du gaz. Seuls 2% de ces jours présentent un niveau moyen de consommation d'électricité. Ces situations de « petits pourcentages » sont répétées dans plusieurs cas

comme le montre la figure 5. Nous avons considéré de façon arbitraire que tous les cas de proportion inférieure à 20%, sont considérés comme des cas des jours de fonctionnement rare et les autres cas sont des jours de fonctionnement fréquent. Ce dernier critère devrait être défini pour chaque branche de l'arbre en accord avec le gestionnaire du bâtiment.

Quand un jour est classé dans la catégorie « rare », une alarme pourra être déclenchée et la personne en charge du suivi pourra analyser plus finement les évolutions des consommations correspondantes et le cas échéant prendre les décisions s'imposant.

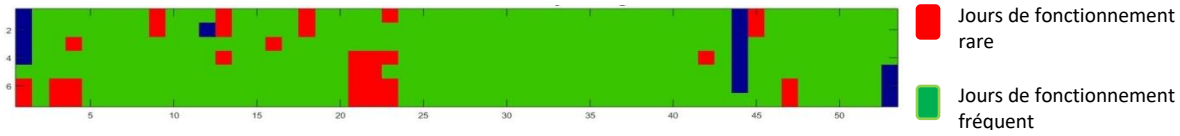

Figure 6 : Jours de fonctionnement rares et fréquents au collège B 2015 le matin

La figure 6 permet de repérer sur la représentation en calendrier les jours correspondant à une situation rare de combinaison pour l'année 2015.

Un autre arbre a été construit pour l'année 2016 dans le but de confirmer les répartitions des combinaisons obtenues en 2015. Les deux arbres ont donné des proportions similaires pour les branches. La très bonne concordance qui a été constatée permet de valider le modèle statistique obtenu. Au final, parmi 27 ($=3^3$) combinaisons possibles des 3 pôles de consommations, 15 sont considérés comme cas rares et 12 comme cas fréquents.

4. CONCLUSIONS ET PERSPECTIVES

Notre premier objectif dans la démarche d'exploitation des données issues de relevés de consommation dans un établissement scolaire était de proposer une procédure automatisable de détection de jours de fonctionnement rare en étudiant les combinaisons de consommation. Pour réaliser cette mission, les jours ont été regroupés tout d'abord dans des familles selon leurs ressemblances de niveaux de consommation des 3 pôles par K-Means. Ensuite, un arbre de décision a combiné les résultats obtenus sur les classifications des 3 pôles de consommation, et des proportions pour chaque catégorie de fonctionnement des jours ont été trouvées. Enfin les jours rares et les jours fréquents ont été définis par un critère spécifique. Une fois les jours rares détectés, l'expert gestionnaire des flux peut analyser la situation plus finement et éventuellement mettre en place des actions pour corriger le problème si une anomalie est détectée. L'association d'une classification et de la constitution d'un arbre de décision a permis de répondre à cette première attente.

La suite du travail qui est en cours vise à utiliser ce modèle statistique pour analyser les consommations des années suivantes dans le but de :

- Classifier les jours de 2017 selon leurs consommations quotidiennes en se basant sur les classes trouvées dans les deux années précédentes.

- Utiliser le modèle pour traiter le jour même les données issues de la télérelève et générer une alerte automatisée.

5. REMERCIEMENTS

Ce travail a été réalisé dans le cadre d'une collaboration avec le Conseil Départemental du Pas-de-Calais qui prévoit une mise à disposition les données de télé-relève. Les auteurs tiennent en particulier à remercier le Service Innovation Energie du département pour ses contributions.

6. BIBLIOGRAPHIQUES

Capozzoli, Alfonso, Fiorella Lauro, et Imran Khan. 2015. « Fault detection analysis using data mining techniques for a cluster of smart office buildings ». *Expert Systems with Applications* 42 (9): 4324-38. https://doi.org/10.1016/j.eswa.2015.01.010.

Chicco, G., R. Napoli, F. Piglione, P. Postolache, M. Scutariu, et C. Toader. 2004. « Load pattern-based classification of electricity customers ». *IEEE Transactions on Power Systems* 19 (2): 1232-39. https://doi.org/10.1109/TPWRS.2004.826810.

Davies, D. L., et D. W. Bouldin. 1979. « A Cluster Separation Measure ». *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1 (2): 224-27. https://doi.org/10.1109/TPAMI.1979.4766909.

Morbitzer, Christoph, Paul Strachan, et Catherine Simpson. 2004. « Data mining analysis of building simulation performance data ». *Building Services Engineering Research and Technology* 25 (août): 253-67. https://doi.org/10.1191/0143624404bt098oa.

Otukei, J. R., et T. Blaschke. 2010. « Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms ». *International Journal of Applied Earth Observation and Geoinformation*, Supplement Issue on « Remote Sensing for Africa – A Special Collection from the African Association for Remote Sensing of the Environment (AARSE) », 12 (février): S27-31. https://doi.org/10.1016/j.jag.2009.11.002.

Usman, Ghousia, Usman Ahmad, et Mudassar Ahmad. 2013. « Improved K-Means Clustering Algorithm by Getting Initial Cenroids », 9.

Xiao, Fu, et Cheng Fan. 2014. « Data mining in building automation system for improving building operational performance ». *Energy and Buildings* 75 (juin): 109-18. https://doi.org/10.1016/j.enbuild.2014.02.005.

Yu, Zhun, Fariborz Haghighat, Benjamin C.M. Fung, et Hiroshi Yoshino. 2010. « A Decision Tree Method for Building Energy Demand Modeling ». *Energy and Buildings* 42 (10): 1637-46. https://doi.org/10.1016/j.enbuild.2010.04.006.