Fouille de données pour l'analyse du comportement complexe de systèmes photovoltaïque-thermiques en vrai grandeur et in situ intégrés aux bâtiments

Leon GAILLARD[1,2*], Guillaume RUEDIN[1], Stéphanie GIROUX-JULIEN[1], Marc PLANTEVIT[3], Mehdi KAYTOUE[4], Christophe MENEZO[1,2], Jean-François BOULICAUT[4]

[1] Université de Lyon, CNRS, INSA-Lyon, UCBL, CETHIL, UMR5008
F-69621, Villeurbanne, France
[2] Chaire INSA de Lyon / EDF « Habitats et Innovations Energétiques »
Villeurbanne, France
[3] Université de Lyon, CNRS, Université Lyon 1, LIRIS, UMR5205
F-69622, Villeurbanne, France
[4] Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205
F-69621, Villeurbanne, France
[*](auteur correspondant : leon.gaillard@insa-lyon.fr)

Résumé – Dans le contexte des bâtiments et des systèmes innovants à très haute performances énergétiques leur monitoring en condition réelles de fonctionnement devient quasiment indispensable afin de pouvoir bénéficier d'un contrôler les performances réelle et de bénéficier d'un retour d'expérience. Comme le suivi d'une installation s'étend sur des périodes longues afin de mesurer la tenue ou la dégradation des performances et qu'un un nombre important de capteurs est impliqué, il est intéressant de considérer les méthodes dites de « fouille de données » adaptées aux grands volumes de données.

Comme cas d'étude, objet de nos recherches sur les systèmes solaires, nous nous intéressons aux analyses des données issues des prototypes du projet RESSOURCES soutenu par l'ADEME. Les techniques d'analyse ont été développées en interdisciplinarité entre les domaines du traitement de data de masse (Big Data) et de l'énergétique des systèmes solaires. Les composants étudiés sont de type double-peau photovoltaïque (PV) ventilée et font partie de ces innovations face au modèle réglementaire actuellement fondé sur une enveloppe « statique » isolée et étanche. De même que pour le rafraichissement des composants photovoltaïques (PV), la lame d'air située derrière la peau PV peut contribuer à la performance thermique du bâtiment : en été servant de barrière thermique avec un écoulement régi par la convection naturelle (effet cheminée), et en hiver comme dispositif de préchauffage à partir d'une ventilation mécanique pour la récupération de chaleur. En intégrant plusieurs fonctions ainsi que les critères esthétiques, les géométries des composants double-peau PV différent d'un bâtiment à l'autre. Tenant en plus compte de la complexité du milieu urbain, des phénomènes physiques inter-reliés régissant leur comportement, la prédiction et l'évaluation des performances réelles de cette variété de configurations constituent un véritable défi difficile à surmonter. Il est donc essentiel d'exploiter de façon ordonnancée les résultats expérimentaux issus des systèmes PV intégrés au bâtiment en vraie grandeur et in situ pour estimer l'apport de tels concepts et pour élaborer des modèles numériques fiables et robustes permettant de prédire leurs performances (thermique et électrique). Cet article concerne l'analyse du comportement des 3 premiers prototypes du projet, opérationnels depuis 2012 et fonctionnant en mode de ventilation naturelle.

La classification de données en termes des conditions environnementales ou de comportement du système est réalisée à partir d'algorithmes tels que le clustering, et l'arbre de décision. Cette approche permet de distinguer les périodes de comportement normal, les défaillances et les défauts d'instrumentation.

Nomenclature

G	rayonnement solaire, Wm^{-2}
k	nombre de clusters
P_c	puissance crête du champ PV, *W*
P_{dc}	production électrique instantanée, *W*
Symboles grecs	
η	rapport de performance électrique, -

Indices et exposants	
dir	direct (pyrhéliomètre)
i	incident
iT	horizontal global (toiture)
ref	référence (ex. 1000 W/m²)
1-3	bloc 1 – bloc 3

1. Introduction

L'Europe s'est fixé de diviser les émissions de gaz à effet de serre par un facteur 4 d'ici 2050 par rapport à 1990, et une réduction de 20% d'ici 2020 tout en augmentant de 20% la part des énergies renouvelables. Le secteur du bâtiment y contribue pour une grande partie et des améliorations dans ce domaine peuvent donc contribuer de manière significative aux objectifs nationaux et européens. En France, la réglementation prévoit dès 2020 que les nouveaux bâtiments soient à énergie positive (BEPOS). Les systèmes Photovoltaïques (PV) intégrés (PVIB) peuvent produire de d'électricité localement et contribuer à la diminution des charges thermiques des bâtiments par convection naturelle ou forcée suivant l'intégration. Quant à l'échauffement inhérent au composant PV cristallins dégradant leur rendement électrique et accélérant leur vieillissement, il peut être lui aussi limité par ventilation naturelle ou forcé (les composants sont alors hybrides Photovoltaïques/Thermiques). La performance et le maintien des performances de ces systèmes PV intégrés est en effet très sensible au choix de conception, et puisque la demande pour les systèmes photovoltaïques entièrement intégrés au cadre bâti (façade, toiture) est encore assez récente, peu d'études sont actuellement disponibles sur le sujet.

Pour le déploiement en masse de systèmes PV intégrés aux enveloppes de bâtiments, certains verrous doivent être surmontés. Il existe d'abord le défi de la complexité du milieu urbain, à ce jour méconnu en termes du potentiel de rayonnement solaire direct et diffus et la circulation d'air. Hétérogènes et fluctuantes, les conditions environnementales sont pénalisantes du point de vue de la production énergétique de systèmes classiques [1]. La complexité éventuelle de l'enveloppe PV est également un enjeu majeur : de tels composants sont souvent multifonctions, et doivent répondre à des critères de performance électrique et thermique, mais aussi d'éclairement naturel, d'isolation acoustique, et d'esthétique. Dans ce contexte de complexité, l'expérimentation à l'échelle 1 en conditions réelles de fonctionnement est essentielle pour valider les concepts envisagés et prédire l'évolution (dégradation) des performances [2]. Dans le cadre du projet RESSOURCES, des composants d'enveloppes PV ont été conçus de manière à atteindre une véritable intégration esthétique, structurale, et énergétique. Trois composants d'enveloppes ventilés ont été intégrés sur des bâtiments réels et instrumentés en configuration naturellement ventilée: deux pour les maisons individuelles et l'autre pour un immeuble de bureaux. Cet article concerne le composant de façade présenté sur la figure 1. Le prototype a été installé chez HBS-Technal à Toulouse, sur un bâtiment de bureaux (open space) occupé sur trois niveaux. Le prototype vertical de hauteur 7.7 m, largeur 4.5 m couvre entièrement les deux premiers étages. La largeur de la lame d'air est de 60 cm et la puissance crête de 1.2 kWc. Ce prototype est orienté à l'ouest et les panneaux PV font un angle de 41° par rapport au sud d'où la forme plissée de la façade. Les modules PV sont répartis en 3 champs, les noms « bloc » 1 à 3 correspondant respectivement au champ du haut, du milieu et du bas. Chaque bloc est relié à une charge résistive constante permettant de suivre la puissance produite.

Des études précédentes [3] ont montré qu'en dépit de la complexité géométrique des prototypes et de leur environnement il ressort une certaine régularité et périodicité de leur comportement sur les échelles temporelles d'une journée ou une année. Cependant, les données issues d'installations PV monitorées contiennent une richesse d'information bien plus importante mais difficilement exploitable : suivi sur plusieurs années avec des pas de temps d'acquisition ~1 minute impliquant des grandes masses de données à traiter ; grand nombre de variables intercorrélés ; multiplicité des échelles spatiales et temporelles en partie dues aux sollicitations environnementales ; limite de résolution spatiale (discrétisation des mesures) insuffisante pour capter certains phénomènes physiques tels que l'effet d'ombrage et du vent. Afin de compléter les techniques conventionnelles utilisées et pour poursuivre l'exploitation de ce suivi, nous nous sommes orientés vers des méthodes adaptées au traitement de grandes masses de données. Cette approche multidisciplinaire est basée sur les sciences de données. Les travaux présentés rentrent dans le cadre d'une collaboration des laboratoires CETHIL et LIRIS qui a débuté en 2012 à travers une contribution au défi CNRS MASTODONS, mission interdisciplinaire du CNRS et poursuivi par le BQR INSA SOLSTICE. Les efforts sont actuellement ciblés sur l'application de méthodes dites « fouilles de données » bien établies ainsi que la création d'outils d'exploration de données. Peu usitées dans le domaine de l'énergie solaire en milieu urbain, ces méthodes montrent leur pertinence dans le domaine de la biologie (expression de gènes) depuis plusieurs années [4].

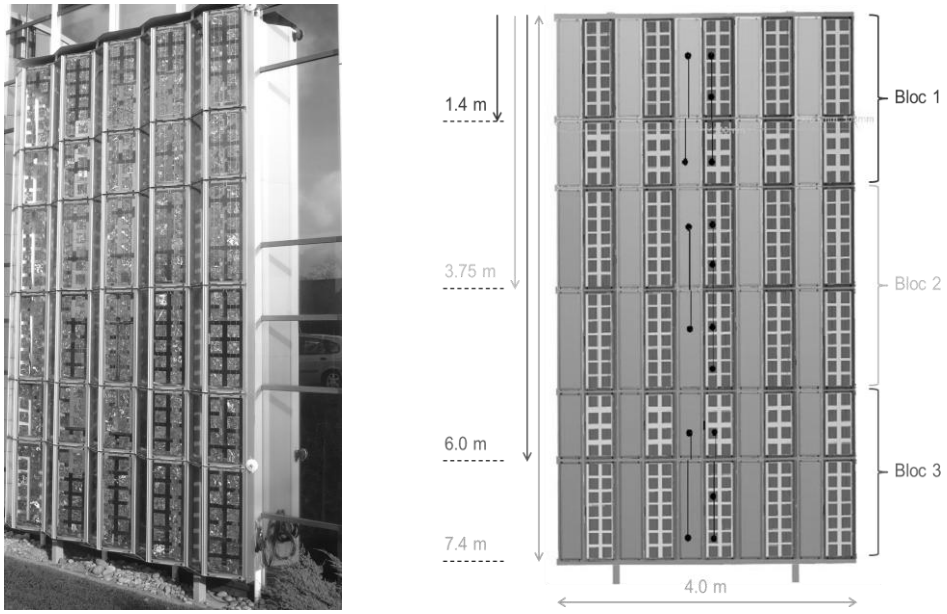

Figure 1 : *Prototype double-peau PV-T du projet RESSOURCES installé à HBS-Technal. A droite : Schéma du prototype indiquant sa répartition en trois blocs.*

2. Méthodes de fouilles de données

L'analyse a été élaborée à partir du logiciel Knime [5]. Une chaine d'analyse, comprenant l'interface avec la base de données MySQL, l'agrégation spatiale et temporelle, le filtrage de données a tout d'abord été implémenté sur un ensemble de données capitalisées sur un an. Plusieurs algorithmes de fouille disponibles dans la bibliothèque standard de Knime ont alors été appliqués sur ce jeu de données prétraitées, et notamment les algorithmes k-means (clustering) et arbres de décision. La méthode de clustering k-means [6,7] est à ce jour éprouvée et l'une des plus utilisées. Cet algorithme de partitionnement de données permet de

diviser un échantillon en k clusters (partitions) attachés à une valeur centrale (ou centroïde). Les clusters sont définis en termes des dimensions, c'est-à-dire un sous-ensemble de variables influentes choisi parmi les variables disponibles. Pour le cas d'application actuel il s'agit de configurer l'algorithme k-means avec les mesures de l'état du système et de son environnement, telles que les températures, intensités de rayonnement incident ou indices de performances. En faisant cela, les clusters obtenus correspondent à une répartition des données en fonction de ces mesures / variables. La solution est trouvée de manière itérative, commençant d'abord par la génération aléatoire de k centres définissant les clusters. Chaque point de donnée est alors associé au centre le plus proche. La répartition réalisée, on détermine les centroïdes (métrique euclidienne) des clusters suivi d'une nouvelle association des points autour de ces centroïdes. Le calcul des centroïdes des k partitions et l'association des données sont alors répétés jusqu'à l'association de clusters cesse d'évoluer (ou un nombre d'itération maximale est atteint). Les clusters obtenus fournissent des informations complémentaires pour l'analyse des données, proposant une description des données par groupes de caractéristiques similaires, telles que les périodes de comportement normal ou singulier. Un point fort de cet algorithme est sa rapidité, et sa capacité de réaliser une analyse multidimensionnelle des grandes masses de données. Les principales faiblesses de la méthode k-means sont les variables et le nombre de partitions à étudier ainsi qu'une certaine sensibilité au point de départ du calcul (effet de 2ème ordre). Ce la nécessite alors de lancer l'algorithme plusieurs fois sur des configurations différentes afin d'évaluer l'impact des choix de l'utilisateur permettant de trouver une solution exploitable. Au delà des centroïdes et dispersions des clusters, l'algorithme k-means ne fournit pas d'information concernant l'appartenance des points à un cluster donné. Il est donc nécessaire d'évoquer d'autres méthodes pour analyser les clusters obtenus. Les algorithmes de type arbre de décision, et plus particulièrement de type arbre de classification, fournissent une description des données réparties à priori en classes (telles que les partitions obtenues par l'algorithme k-means) [8]. Pour distinguer des classes, une série de tests basée sur la logique booléenne est effectuée, fournissant ainsi un schéma arborescent avec deux « sous-populations » attachées à chaque nœud. Le but de l'algorithme de classification est de trouver le seuil optimal définissant les deux « sous-populations » les plus distinctes. Ce seuil optimal est évalué par une mesure de qualité, le facteur Gini étant les plus souvent utilisé [9].

3. Clustering de la performance électrique et du rayonnement solaire

Nous nous sommes particulièrement intéressés aux relations de dépendance de la production électrique. Comme il a été précédemment rapporté [3], ces corrélations présentent différentes tendances qui résultent d'une superposition de comportements distincts. L'algorithme de clustering a été utilisé afin d'identifier et d'isoler ces différentes tendances. La figure 2 présente la meilleure répartition obtenue, pour k=6, où les dimensions retenues comprennent le rayonnement solaire (dans le plan de la façade, le global sur un plan horizontal, et direct), ainsi que le coefficient de performance de chaque bloc défini par l'équation 1. L'appartenance de chaque point de données à un cluster est indiquée par le style de marqueur. Le numéro de marqueur étant arbitraire, les clusters consécutifs ne sont pas plus similaires qu'aux autres. Cependant, la liste de clusters est présentée dans la légende de chaque graphe en ordre de population décroissante.

$$\eta = \left(\frac{P_{dc}}{P_c} \right) \Big/ \left(\frac{G_i}{G_{ref}} \right) \quad (1)$$

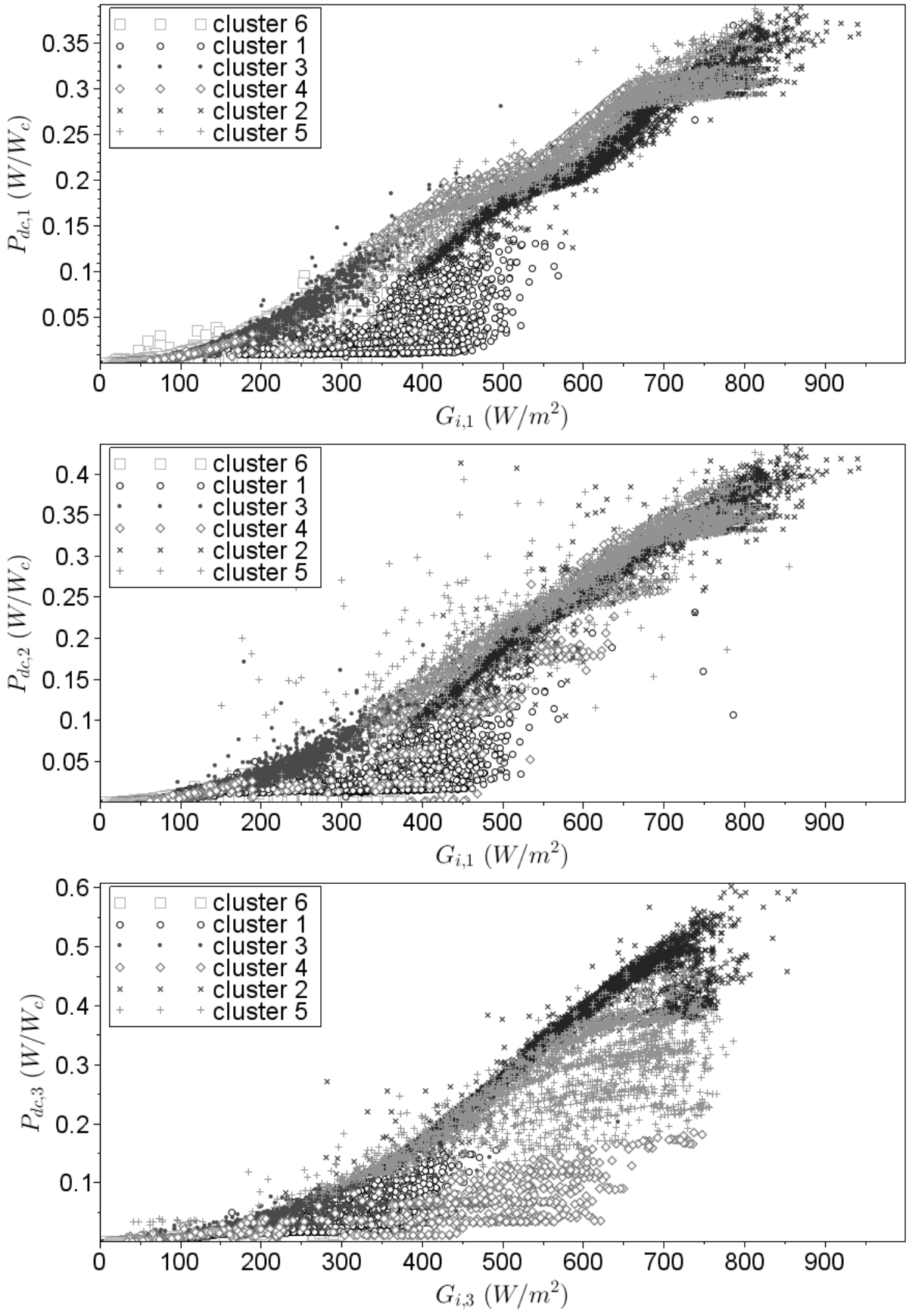

Figure 2 : *Corrélations de rayonnement solaire incident et de production électrique par bloc, avec une répartition en clusters indiquée par le style de marqueurs. Du haut au bas : bloc 1 à bloc 3.*

Comme le montre la figure 2, certains clusters sont localisés sur le graphe. Par exemple, le cluster 6 est circonscrit sur les faibles éclairements et faibles performances pour les trois graphes. De même, on constate que le cluster 2 est associé aux périodes de forte performance pour les trois blocs. En revanche, pour un rayonnement solaire assez important, le cluster 4 est caractérisé par une performance élevée pour le bloc 1, une performance réduite du bloc 2, et une faible performance du bloc 3. Concernant la dépendance production électrique / rayonnement global sur un plan horizontal, une visualisation de type « carpet » des clusters trouvés (k=6) pour le bloc 3 (partie inférieure de la façade est présentée figure 3. Sur les graphes, les valeurs manquantes dues principalement à des interruptions ponctuelles d'acquisition sont remplacées par des valeurs lissées sur trois points consécutifs. La distribution de rayonnement est caractérisée par une enveloppe saisonnière sinusoïdale, les variations ponctuelles étant dues à la nébulosité. En regard, les données relatives à la performance électrique affichent des structures horizontales correspondant à des chutes de production à la même heure sur plusieurs semaines dues à la présence d'obstacles fixes.

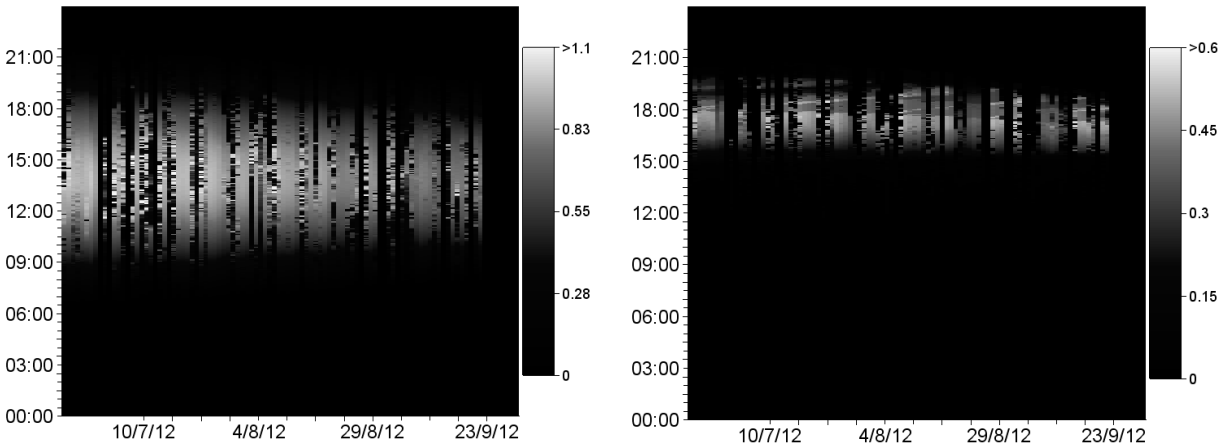

Figure 3 : *Visualisation du rayonnement solaire global à l'horizontal (gauche) et de la production électrique du champ bloc 3 (droite) sur les axes de date et heure. L'échelle de couleurs donne l'intensité normalisée à 1000W/m² et à la puissance crête de l'installation respectivement.*

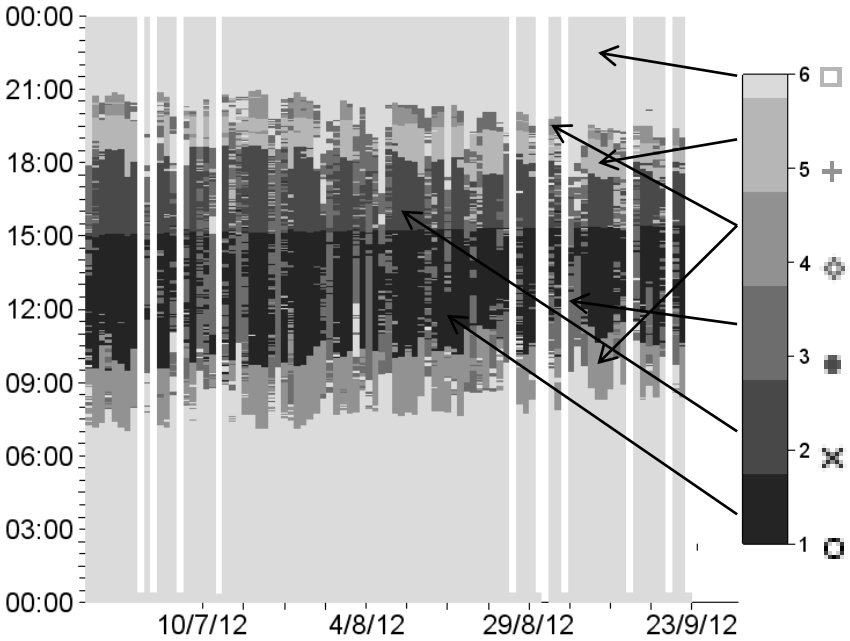

Figure 4 : *Visualisation de clusters obtenus par algorithme k-means sur les axes de date et heure.*

La même classe de visualisation a été utilisée pour étudier les clusters sur la figure 4. Il s'avère que plusieurs clusters sont corrélés avec l'heure. Cette représentation des résultats sert ainsi à compléter l'interprétation initiale des clusters affichée sur les graphes de corrélations. Les résultats correspondant à la nuit sont associés au cluster 6. De même, le cluster 4 regroupe les périodes de l'aube et du crépuscule. En confrontant les visualisations « carpet » de la figure 3, le cluster 3 est évidemment associé aux périodes très nuageuses. Pour les jours partiellement nuageux, la production électrique se réparti en plusieurs clusters indicateur d'un comportement complexe et connexe. Le reste des résultats sont principalement repartis en 3 clusters corrélés avec les bandes horizontales visibles dans la figure 3.

4. Classification de clusters par arbre de décision

Suite à la répartition en clusters, une analyse par arbre de décision a été effectuée afin d'obtenir une description des clusters en termes de facteurs les plus influents. Les résultats sont résumés dans la figure 5. Malgré l'introduction d'un nombre de paramètres plus important que les dimensions choisies pour l'algorithme k-means, l'arbre de décision résultant se compose finalement des règles basées sur le rayonnement global sur un plan horizontal, le rayonnement direct mesuré par pyrhéliomètre, et le coefficient de performance instantanée du bloc 3. Une série de 4 branchements permet de distinguer les 6 clusters avec une efficacité supérieure à 90% pour tous les clusters excepté les numéros 4 et 5.

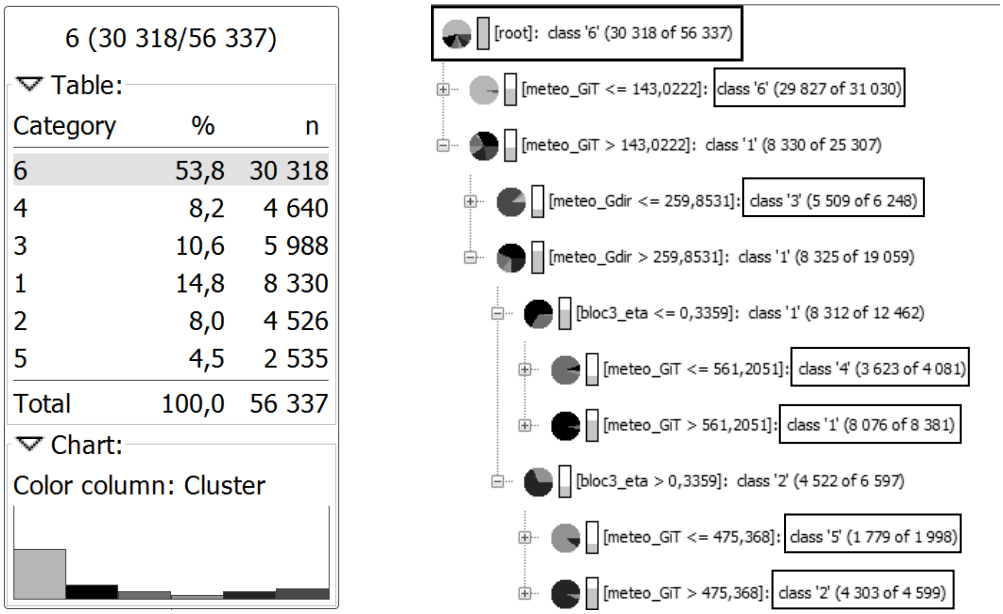

Figure 5 : *Classification de clusters par méthode d'arbre de décision*

En intégrant les informations acquises par les visualisations des résultats enrichis par l'arbre de décision, les clusters ont été interprétés en termes des conditions environnementales (voir le tableau 1). Par la suite il serait possible d'exploiter les règles de classification comme des filtres efficaces pour traiter d'autres séquences de données issues de la même installation, ou d'étudier plus précisément les partitions de données déjà obtenues (sous-clustering). Bien que l'utilité de la méthode combinée de visualisation, clustering et arbre de décision ait été démontrée pour la classification semi-automatique des données selon les caractéristiques apparentes, les résultats seuls de cette analyse n'induisent pas une identification des phénomènes physiques pilotant des changements de comportement du système. Afin de faire émerger des connaissances concernant la réponse du système assujetti aux sollicitations externes, il sera nécessaire d'approfondir et de confronter les données traitées avec des modèles, soit physiques, soit comportementaux.

Cluster	*Classification*	*Interprétation*
1 (97%)	(G_{iT}>561 | (η_3<=33,5% | (G_{dir}>260 | (G_{iT}>143))))	Ensoleillé, diffus sur façade
2 (95%)	(G_{iT}>475 | (η_3>33,5% | (G_{dir}>260 | (G_{iT}>143))))	Façade pleinement éclairé
3 (92%)	(G_{dir}<=260 | (G_{iT}>143))	Nuageux
4 (78%)	(G_{iT}<=561 | (η_3<=33,5% | (G_{dir}>260 | (G_{iT}>143))))	Bloc3 à l'ombre
5 (70%)	(G_{iT}<=475 | (η_3>33,5% | (G_{dir}>260 | (G_{iT}>143))))	Ombrage partiel
6 (98%)	(G_{iT}<=143)	Nuit

Tableau 1 : *Classification et interprétation des clusters par arbre de décision. Les pourcentages indiquent la qualité de la règle de classification*

5. Conclusions

Nous avons introduit l'utilisation des méthodes dites « fouilles de données » dans le cadre d'une étude d'évaluation expérimentale des installations PV déployées en zones urbaines. Les méthodes servent à prétraiter les résultats et contribuent à l'interprétation des comportements constatés. Afin de démontrer l'utilité de ces méthodes, les données du projet RESSOURCES ont constitué un cas d'étude. Les algorithmes de type clustering k-means et arbre de décision (classification) ont été exploités afin de distinguer de manière semi-automatique des conditions environnementales différentes et des périodes de comportements dit normaux ou singuliers. Concernant les méthodes de clustering, il est important à ce stade de systématiser le choix des dimensions (paramètres) et de mettre au point une méthode permettant d'évaluer et comparer des résultats sur une gamme de nombre k.

Références

[1] M. A. Eltawil, Z. Zhao, Grid-connected photovoltaic power systems : Technical and potential problems – a review, *Renewable and Sustainable Energy Reviews* 14 (2010) 112-129

[2] M.D. Bazilian, F. Leenders, B.G. Van Der Ree, D. Prasad, Photovoltaic cogeneration in the built environment, *Solar Energy* 71 1 (2001) 57-69

[3] L. Gaillard, S. Giroux-Julien, C. Ménézo, H. Pabiou, Experimental evaluation of a naturally ventilated PV double-skin envelope in real operating conditions, *Sol. Energy* 103 (2014) 223-241

[4] H. Hackl et al., Molecular processes during fat cell development revealed by gene expression profiling and functional annotation, *Genome Biology* 6 (2005) 13, R108)

[5] M. R. Berthold et al., KNIME: The Konstanz Information Miner, *Data Analysis, Machine Learning and Applications*, Springer Berlin Heidelberg (2008), pp319-326

[6] H. Steinhaus, Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci.* IV (C1.III) (1956) 801–804

[7] K. J. Anil, Data clustering: 50 years beyond K-means, *Pattern Recognition Letters*, 31 (2010) 8 651-666

[8] J. Shafer, R. Agrawal, M. Mehta. SPRINT: A scalable parallel classifier for data mining. *Proc. 2nd Int'l Conf. on Very Large Databases*, (Bombay, India, September 1996)

[9] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone. *Classification and Regression Trees*, Wadsworth, Belmont, 1984

Remerciements

Ce travail est soutenu par les projets RESSOURCES (Convention ADEME 0705C0076), BQR SOLSTICE (INSA-Lyon), et CNRS AMADOUER (défi MASTODONS) ainsi que par la Chaire INSA/EDF « Habitats et innovations énergétiques ».