Mesures de similarité pour l'aide à l'analyse des données énergétiques de bâtiments

Hala Najmeddine, Frédéric Suard, Arnaud Jay, Philippe Marechal, Marié Sylvain

Mesures de similarité pour l'aide à l'analyse des données énergétiques de bâtiments

Hala Najmeddine [1] Frederic Suard [1] Arnaud Jay[2] Philippe Maréchal [2] Sylvain Marié [3]
[1] CEA, LIST, 91191 Gif-sur-Yvette CEDEX prenom.nom@cea.fr
[2] CEA, INES, Savoie-Technolac, 73377 Le Bourget-du-Lac CEDEX prenom.nom@cea.fr
[3] Schneider Electric, Energy Efficiency Innovation, Site 38TEC,37 Quai Paul-Louis Merlin, 38050 Grenoble Cedex 9 prenom.nom@schneider-electric.com

Résumé

Cet article propose de comparer différentes métriques de séries temporelles, afin de suggérer les méthodes les plus adaptées pour l'analyse de données énergétiques du bâtiment. Dans un premier temps, la comparaison porte sur les métriques de signaux monodimensionnels, afin d'établir des préconisations selon l'objectif de l'analyse et la nature des informations. Dans un deuxième temps, afin de résoudre la problématique d'analyse sur des grandes bases de capteurs, une métrique supplémentaire est présentée permettant de comparer des signaux multidimensionnels. Ces différentes métriques sont appliquées sur des données réelles issues de bâtiments démonstrateurs équipés de plusieurs centaines de capteurs fournissant des mesures en continu. L'objectif initial portant sur l'aide au diagnostic, les métriques sont comparées grâce aux résultats fournis par un algorithme de projection de données : Isomap, afin de proposer une méthode complète de traitement et d'analyse. Les résultats mettent en avant les différences entre ces différentes métriques et l'intérêt d'utiliser une métrique multidimensionnelle pour faciliter l'exploitation des données.

Mots Clefs

Séries temporelles, fouille de données, mesure de similarité, capteur, diagnostic et aide à la décision, INCAS

Abstract

The classification and clustering methods of time series have been effective in providing useful information in various fields. The past researchs on similarity between temporal databases have been the subject of numerous scientific studies. Many methods of classification and similarity measures have been developed over the years, unfortunately most of these measures can not be used directly on time series. This paper summarizes the basis of similarity between time series, including similarity measures and visualisation algorithms for one dimension and multidimensional data. These similarity measures are computed on real data from experimental buildings of the INCAS platform equipped with hundreds of sensors. Moreover, analyzing energy data will help researchers to improve modeling and prediction of energy consumption in smart buildings.

Keywords

Time series, Data mining, similarity measures, sensors, diagnosis and decision support, INCAS.

1 Introduction

L'optimisation des performances énergétiques et du fonctionnement des systèmes thermiques et électriques des bâtiments [1] nécessite une instrumentation de plus en plus importante afin de comprendre et d'analyser les différents phénomènes permettant d'appréhender au mieux l'approche globale de l'efficacité énergétique du bâtiment [2]. C'est un problème complexe qui comprend de très nombreuses incertitudes ainsi que des paramètres plus ou moins bien maîtrisés par les acteurs du domaine : impact des occupants, conditions extérieures, phénomènes météorologiques, la qualité de l'air, le confort thermique, etc.

Pour répondre à cette problématique, l'Institut National de l'Énergie Solaire – INES – a construit des maisons expérimentales proches du standard passif depuis 2009 [3]. Ces maisons expérimentales, totalement instrumentées, possèdent diverses structures d'enveloppe avec plus ou moins d'inertie. Ces maisons sont équipées avec une centaine de capteurs mesurant des informations variées telles que la température, l'humidité ou encore la vitesse d'air.

Cependant, l'analyse et l'extraction des phénomènes physiques implique de mettre au point des méthodes d'analyse automatique capables de comparer ces différents signaux en vue de proposer une solution de supervision et d'analyse visuelle des données disponibles. Les outils d'analyse sont dédiées aux professionnels du bâtiment afin de permettre l'intégration de la gestion énergétique du bâtiment dès les phases de conception.

Nous proposons donc dans cet article une méthode pour l'aide à l'analyse des données énergétiques du bâtiments. L'objectif principal consiste à mettre au point un traitement intégrant des outils de fouille de données, afin de proposer aux thermodynamiciens des outils capables d'appréhender le volume de données disponibles et optimiser ainsi l'étude du bâtiments. Il s'agit ainsi de compléter des premiers travaux orientés sur la classification de données [4]. Dans le

cadre de ces travaux, il est apparu que la qualité de la classification de séries temporelles était fortement impactée par la mesure de similarité. Ainsi, la plupart des métriques fondées uniquement sur la distance euclidienne [5] ne sont pas robustes aux problèmes de déphasage ou d'amplification qui sont très courant dans le domaine à cause de l'inertie des murs notamment.
Nous proposons ici d'utiliser une amélioration de la distance euclidienne introduite par Keogh et al. qui intègre la complexité globale des signaux (CID [6]). L'état de l'art de ce domaine établi par Liao et al. en 2005 [7] montre que d'autres approches existent pour s'affranchir des limitations de la distance euclidienne, notamment la DTW [8] qui permet de mettre en correspondance des signaux instantanément différents. Cependant, la formulation originale de la DTW souffre de quelques limitations. Nous présentons ici des améliorations de la DTW (DDTW [9] et AFBDTW [10]). Enfin, nous introduisons une mesure de similarité sur des ensembles de signaux, afin de réduire la mesure de similarité multidimensionnelle *Eros* [11].
Afin de pouvoir faciliter l'interprétation de ces mesures de similarité plusieurs approches sont possibles [12], telles que la classification hiérarchique [4], le clustering k-mean, etc. Nous proposons ici d'appliquer des méthodes de réduction de dimension, afin de projeter dans un espace restreint, donc visualisable, en utilisant la matrice de distance entre les données. Nous utilisons l'algorithme Isomap [13], qui fait partie de l'état de l'art. En effet, il nécessite peu de paramètres et sa complexité reste faible par rapport au nombre de données exploitées, ce qui le destine tout particulièrement à l'application interactive qui est l'objectif à plus long terme.
Cet article présente dans un premier temps les métriques utilisées pour la comparaison de signaux simples, avant de détailler dans un deuxième temps la métrique proposée pour les signaux multidimensionnels. Enfin, la troisième partie présente les résultats obtenus sur les données des bâtiments démonstrateurs et établit un comparatif selon la nature des signaux et les objectifs de diagnostic.

2 Mesure de similarité

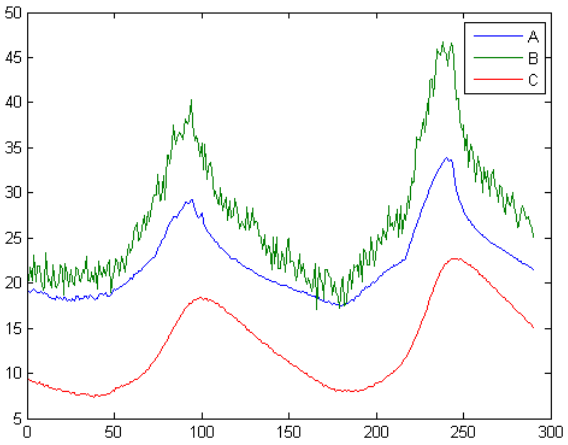
Plusieurs méthodes proposent de mesurer le degré de similitude entre des séries temporelles monodimensionnelles, à commencer par le point de départ, la distance euclidienne "**Euclidian Distance (ED)**" [5, 14]. Pour deux vecteurs v et u de taille N, la distance ED est exprimée comme étant :

$$d_{ED}(u,v) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(u_i - v_i)^2} \qquad (1)$$

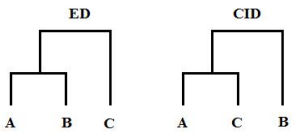
Comme la mesure de similarité ED n'a pas de borne supérieure et que sa valeur augmente avec le nombre de descripteurs N, il est conseillé de calculer la distance ED normalisée. De plus, cette distance ignore les dépendances temporelles entre les différentes séries de données. Ces deux contraintes ne permettent donc pas de comparer la forme du signal, ce qui est contradictoire avec l'objectif de classification dans notre cas. "**The Complexity Invariance Distance (CID)**" introduite par Keogh en 2011 [6], propose une notion plus robuste que la distance euclidienne, car elle comporte un facteur correcteur que nous pouvons associer à une nouvelle mesure de similarité. La distance CID s'écrit comme suit :

$$d_{CID}(u,v) = d_{ED}(u,v) \cdot \frac{max\{CE(u), CE(v)\}}{min\{CE(u), CE(v)\}} \qquad (2)$$

avec $CE(x) = \sqrt{\sum_{i=1}^{N-1}(x_i - x_{i-1})^2}$ l'estimation de la complexité de la série temporelle x. Le facteur de complexité étant calculé en cumulant l'ensemble des variations locales du signal.
La figure 1 illustre les différences entre la distance ED et CID en effectuant une classification de 3 signaux. La classification hiérarchique basée sur la distance ED (cf. Fig. 1 (b) à gauche) a regroupé les signaux A et B de formes similaires tandis que la distance CID (cf. Fig. 1 (b) à droite) a regroupé les signaux A et C de variations semblables.



(a)



(b)

FIGURE 1 – (a) Trois séries temporelles A, B et C. (b) Classification hiérarchique des séries temporelles A, B et C basée sur les distances ED et CID.

Dans le domaine des données énergétiques, de nombreuses études portent sur la comparaison des mesures à différents endroits du bâtiment. A cause du phénomène de propagation de la chaleur, les signaux mesurant une même grandeur physique sont décalés dans le temps. Pour résoudre le problème de distorsion dans les séries temporelles, Sankoff et Kurskal [8, 15] ont présenté la distance de déformation temporelle "**Dynamic Time Warping (DTW)**" qui

considère que le temps est élastique et non pas linéaire. La mesure de similarité sur un ensemble de séries temporelles brutes contenant le phénomène de distorsion avec la distance ED peut induire en erreur. Cette distance est en effet très sensible aux effets de déphasage. Ainsi la distance ED entre les deux séries temporelles de la figure 2 n'est pas négligeable alors qu'elles pourraient être considérées comme deux séries semblables avec la DTW.

La DTW est utilisée dans les travaux de comparaison de séries temporelles de Keogh [16] et Park [17].Comme l'illustre la figure 2, la DTW permet de comparer deux séries temporelles de dimension différente. Le principe de la distance consiste à mettre en correspondance les sous-séquences qui "se ressemblent" même si elles ne correspondent pas à un même intervalle de temps (voir Fig. 2 où les points appariés des deux séquences temporelles contribueront au calcul de la distance DTW).

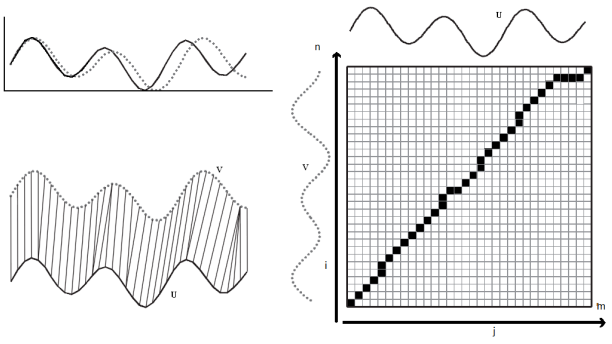
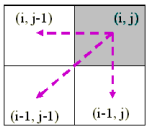


FIGURE 2 – Pour deux séries temporelles, un exemple d'application de la DTW pour la mise en correspondance des points à partir de la matrice de distance cumulée

La comparaison de deux séries temporelles U et V de dimensions respectives m et n, est basée sur la réplication des valeurs jusqu'à l'obtention de la meilleure correspondance. Pour le calcul de la matrice de distance de dimension $(n \times m)$ représentée dans la Fig. 2, on initialise $d_{cum}(1,1) = \mid u_1 - v_1 \mid$. Après initialisation de la première ligne et la première colonne de la matrice $d_{cum}(1,j) = \mid u_j - v_1 \mid + d_{cum}(1,j-1)$ pour $j > 1$ et $d_{cum}(i,1) = \mid u_1 - v_i \mid + d_{cum}(i-1,1)$ pour $i > 1$, les valeurs des autres cases sont calculées comme suit pour $i + j > 2$:

$$d_{cum}(i,j) = d_{ED}(u_i, v_j) + min\{d_{cum}(i-1, j-1)$$

$$d_{cum}(i-1, j), d_{cum}(i, j-1)\}$$



Pour relier les points U et V, il faut trouver le chemin qui minimise la distance cumulée. Ce chemin W est formé de K points avec $max(n,m) \leq K \leq n + m - 1$, k étant le couple (u_i, v_j) sélectionné. La distance DTW optimale est :

$$d_{DTW}(U,V) = min\sqrt{\sum_{k=1}^{K} d_{ED}(k)} \quad (3)$$

Selon ce principe, la DTW a tendance à expliquer les variations de l'axe des Y en déformant l'axe des X. Cela peut cependant induire à des alignements non désirables. Pour répondre à ce problème Keogh et Pazzani ont proposé en 2001 une modification de la DTW nommée **"Derivative Dynamic Time Warping (DDTW)"** [9]. Cette méthode prend en compte la forme des séries temporelles et de la première dérivée des séquences. Le premier terme dans le calcul de la distance cumulée n'est plus celui de la distance ED $d_{ED}(u_i, v_j)$, mais celui de l'estimation de la dérivée de u_i et v_i :

$$d_x(v_i) = \frac{(v_i - v_{i-1}) + (v_{i+1} - v_{i-1})/2}{2} \quad (4)$$

La DDTW fournit des performances nettement supérieures à celle de la DTW originale en minimisant le nombre de points dupliqués [9].

Cependant, ni la DTW, ni la DDTW ne permettent de trouver un alignement dans le cas de données manquantes, ce qui est très fréquent dans le cas de mesures réelles. L'**"Adaptive Feature Based Dynamic Time Warping (AFBDTW)"** présenté en 2010 par Xie et Wiltgen [10] permet de prendre en compte à la fois le caractère local ainsi que global des séries pour les correspondances au lieu de la valeur elle même ou de sa dérivée.

Le caractère local de u_i nommé $f_{local}(i)$ est défini par :

$$f_{local}(i) = (u_i - u_{i-1}, u_i - u_{i+1}) \quad (5)$$

Il semblerait que cette définition représente d'une meilleure façon le caractère global par rapport à la dérivée de la DDTW. Le caractère global est défini par :

$$f_{global}(i) = \left(u_i - \frac{1}{i-1} \sum_{k=1}^{i-1} u_k,\ u_i - \frac{1}{m-i} \sum_{k=i+1}^{m} u_k \right) \quad (6)$$

Pour l'évaluation de la distance entre u_i et v_j, on définit $dist(u_i, v_j)$ comme suit. Le calcul de d_{cum} restera le même que celui défini précédemment.

$$dist(u_i, v_j) = W_1 \cdot dist_{local}(u_i, v_j) + W_2 \cdot dist_{global}(u_i, v_j)$$

avec

$$\begin{aligned} dist_{local}(u_i, v_j) &= \mid (f_{local}(u_i))_1 - (f_{local}(v_j))_1 \mid \\ &+ \mid (f_{local}(u_i))_2 - (f_{local}(v_j))_2 \mid \\ dist_{global}(u_i, v_j) &= \mid (f_{global}(u_i))_1 - (f_{global}(v_j))_1 \mid \\ &+ \mid (f_{global}(u_i))_2 - (f_{global}(v_j))_2 \mid \\ W_1 + W2 = 1, 0 \leq W_1 \leq 1,\ 0 \leq W_2 \leq 1 \end{aligned} \quad (7)$$

Les poids W_1 et W_2 permettent de régler le pourcentage d'influence du critère local et global. La figure 3 présente la

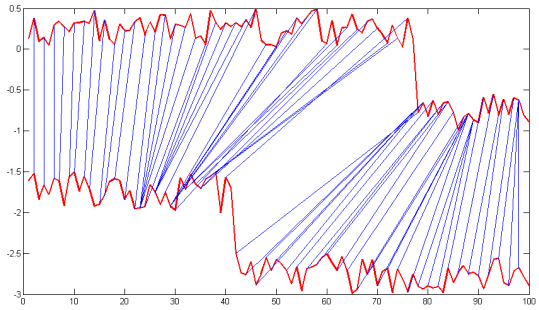


FIGURE 3 – Mise en correspondance avec AFBDTW

mise en correspondance des données de deux séries temporelles présentant une rupture dans le temps. La mise en correspondance des séries temporelles selon un critère local et un critère globale rend la AFBDTW plus complexe que les autres méthodes présentées ci-dessus, mais parfaitement adaptée pour les séries temporelles souffrant de coupures, présentant des variations similaires de fréquences, d'amplitudes de déformations avec un décalage de temps. Le poids reliant deux caractéristiques similaires sera inférieur aux autres poids établis dans la mise en correspondance et donc sera automatiquement sélectionné par le critère de minimisation.

Les trois méthodes basées sur la DTW sont comparées dans la figure 4 où deux couples de séries temporelles sont traitées par la DTW, la DDTW et la AFBDTW. Visuellement, la mise en correspondance issue de la DTW est la moins efficace des trois. Les deux autres méthodes DDTW et AFBDTW ont permis de faire une bonne correspondance et une détection des formes. Notons tout de même que la duplication des points reste raisonnable avec la AFBDTW. Sachant que la mesure de similarité se base sur les nouveaux signaux dupliqués, les distances calculées avec la DTW sont nettement supérieures à celles calculées avec les deux autres méthodes. La classification des séries temporelles basée sur la mesure de similarité avec la AFBDTW, est le meilleur compromis, permettant de prendre en compte à la fois le critère local et global des séries temporelles.
Ces dernières années, de nombreuses méthodes ont ainsi été proposées pour mesurer la similarité entre des données temporelles, chacune présentant des avantages et des faiblesses. Selon le domaine d'application visé et les données considérées, la problématique revient donc à choisir tout d'abord la mesure de similarité la plus adaptée. A travers ces différentes mesures, il est possible de répondre à la plupart des problématiques posées, notamment en terme de séries non-synchronisées et bruitées. Cependant, aucune méthode ci-dessus ne permet de comparer des ensembles de signaux. Une mesure de similarité multidimensionnelle permettrait en effet de réduire la complexité de l'analyse en proposant une classification sur des groupes de signaux.

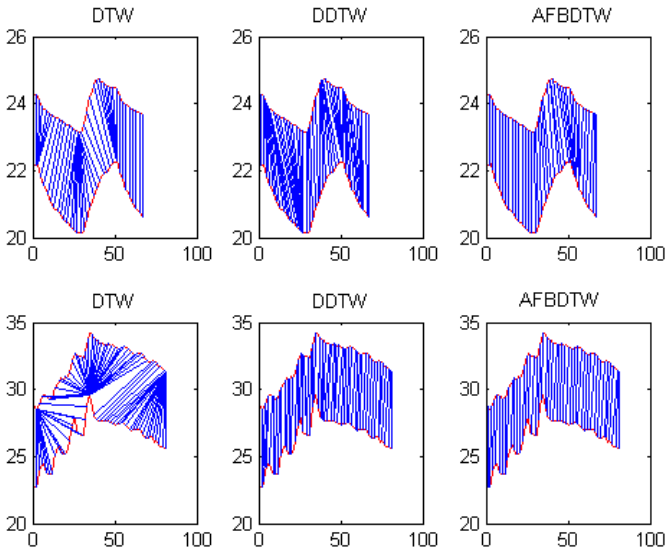


FIGURE 4 – Comparaison visuelle des performances des trois méthodes de mise en correspondance avec la DTW, DDTW et AFBDTW

3 Mesure de similarité multidimensionnelle

La mesure de similarité multidimensionnelle vise à indiquer le niveau de similitude entre plusieurs bases ou groupes de données d'une façon simultanée. Contrairement aux autres méthodes citées qui cherchent le niveau de similarité entre deux variables en omettant la corrélation existante entre l'ensemble des variables, une méthode multidimensionnelle prend en compte la contribution de chacun des variables dans la définition de la mesure de similarité totale des groupes. L'une des méthodes traitant les séries temporelles multidimensionnelles (Multivariate Time Series MTS) est la méthode "$Eros$" (Extended Frobenius norm) [11].
L'intuition derrière cette méthode réside dans sa capacité à traiter un ensemble d'observations (d'individus) différent pour chaque groupe de données avec le même nombre de variables (capteurs) n. Cette méthode s'affranchit du problème de dimension en considérant non pas le groupe de données mais les valeurs propres et vecteurs propres de la matrice de covariance qui, eux, sont de tailles identiques. Un autre avantage apporté par cette méthode concerne la réduction de dimension des données. La taille de la matrice de covariance de chaque base est une matrice $n \times n$, le nombre de données a donc été considérablement réduit puisque le nombre d'observations est généralement supérieur à celui des variables. L'algorithme décrivant le processus de mesure de similarité multidimensionnelle est présenté dans Algorithm 1.
Sachant que A et B sont deux MTS de dimension respective $m_A \times n$ et $m_B \times n$. Considérons V_A et V_B les deux matrices de vecteurs propres dites à droite résultant de la SVD

Algorithm 1 Mesure de similarité en MD

1: Choisir les fenêtres de partitionnement(nombre d'individus) et le nombre de variables n
2: Partitionnement de l'ensemble de données en un ensemble de mini base ayant le même nombre de variables
3: Calculer la matrice de covariance de chaque mini base
4: Décomposer chaque matrice de covariance avec la SVD
5: Récupérer les valeurs propres et les vecteurs propres
6: calculer le poids w des individus en normalisant les valeurs propres [11]
7: Calculer la similarité entre les MTS

(la SVD décompose une matrice M en $U\Sigma V^T$) appliquée sur les matrices de covariance M_A et M_B. V_A et V_B sont exprimées sous la forme suivante : $V_A = [a_1, \cdots, a_n]$ et $V_B = [b_1, \cdots, b_n]$, avec a_i et b_i des vecteurs orthogonaux. La similarité Eros entre les deux MTS A et B est calculée comme suit :

$$Eros\,(A, B, w) = \sum_{i=1}^{n} w_i | < a_i, b_i > | \qquad (8)$$

Or $Eros$ ne satisfait pas l'inégalité triangulaire [18]. Par conséquent, nous définissons D_{Eros}, qui préserve la relation de similitude d'$Eros$ comme suit :

$$D_{Eros}\,(A, B, w) = \sqrt{\left(2 - 2\sum_{i=1}^{n} w_i \left|\sum_{j=1}^{n} a_{ij} \times b_{ij}\right|\right)} \qquad (9)$$

Le tableau 1 présente le coût d'exécution des différentes méthodes de mesure de similarité. Notre étude bibliographique nous a permis d'explorer différentes métriques pour isoler des phénomènes physiques. La distance euclidienne permet de regrouper les séries temporelles d'amplitudes approchées. La CID permet de discriminer les capteurs défectueux et les capteurs de variations semblables. La DTW permet d'identifier les réponses thermiques similaires non synchronisées. Enfin $Eros$ permet de faire une analyse de zone thermique dans le bâtiment. Notre analyse consiste à utiliser la définition de la mesure de similarité entre les séries temporelles traduite par des tableaux de similarités. Cette démarche permettra de séparer la construction de l'algorithme de mesure de similarité entre les données de celui du choix de la représentation. Une proposition de visualisation unique basée sur la projection avec Isomap [13] utilisera toutes les sortes de similitudes. Les experts thermiciens pourront effectuer des comparaison entre les résultats sans se référer à la métrique utilisée.

Eros	ED	DTW
$O\left(m \times n^2 + n^3\right)$	$O\left(m \times n\right)$	$O\left(m^2 \times n\right)$

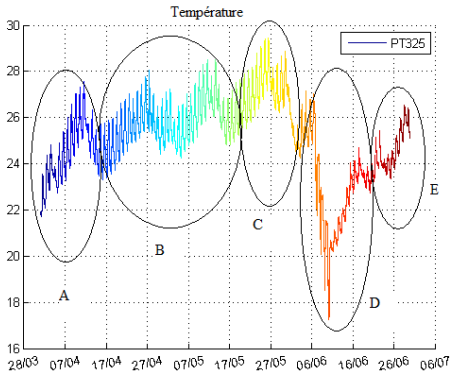
TABLE 1 – Coût des différentes fonctions de mesure de similarité.

4 Application

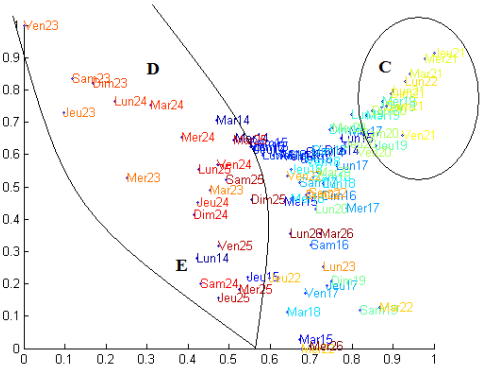
Dans cette partie nous allons présenter les performances expérimentales obtenues en appliquant les différentes méthodes de mesure de similarité. Les deux approches monodimensionnelle et multidimensionnelle seront appliquées sur des données mesurées dans les bâtiments INCAS [3] durant les mois allant de mai à juin 2011. Ces bâtiments sont équipées d'un ensemble de capteurs mesurant à la fois la température dans les différentes pièces, la température extérieure, le rayonnement global, l'humidité relative, la consommation électrique, etc. Ces capteurs sont placés sur les murs, dans l'air et sur les fenêtres. Une mesure est effectuée toutes les minutes. Les détails sur ces constructions et l'emplacement des capteurs sont présentés dans [4].

La figure 5(a) trace l'évolution de la température mesurée par le capteur nommé "PT325" placé dans la cuisine. Le dégradé de couleur passe du bleu au rouge selon l'évolution dans l'axe de temps. Nous avons défini plusieurs zones de fonctionnement pour faciliter la lecture de la mise en correspondance entre les données temporelles et les projections. Ces zones seront par la suite identifiées dans les différentes projections. Trois métriques différentes de mesures de similarité sont comparées. Chaque série temporelle représente les températures mesurées sur une journée. Elle sera représentée avec la même couleur dans les différentes projections et portera le nom de la journée et le numéro de la semaine dans l'année (ex : Lun24 correspond aux données mesurées le 14 juin). La projection des distances calculées avec les trois méthodes ED en Fig. 5(b), CID en Fig. 5(c) et AFBDTW en Fig. 5(d) met en évidence que les semaines 23 et 24 (en rouge et orange représentant les zones D et E avec une sur-ventilation du bâtiment) et les semaines 21 et 22 (en jaune représentant la zone C) contiennent des jours différents sur l'ensemble des données, c'est à dire quand la température a subit des variations importantes. Une seule classe de fonctionnement est identifiée avec les deux métriques ED et CID basées sur un critère local. En revanche, la projection des mesures de similarité fournies par AFBDTW présente deux groupes des journées selon les sous ensembles A et E instables d'une part et l'ensemble des jours de la zone B stable d'autre part. Ces différentes projections facilitent l'interprétation des données temporelles issues des bâtiments démonstrateurs. Elles permettent de décrypter les informations d'une façon indépendante pour chacune des séries temporelles.

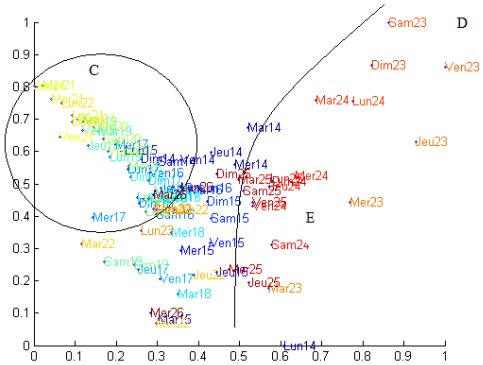
Pour analyser l'évolution journalière d'une zone thermique dans le bâtiment tout en réduisant la dimension des données à visualiser, nous avons utilisé les mesures de similarité calculées avec la distance $Eros$ avec un ensemble représentatif de cinq capteurs. Les données correspondent à la température dans le bâtiment (PTA005, PT325 et THCV005) mesurée à différents endroits, le rayonnement solaire et la température extérieure (cf.Fig. 6(a)). Dans une



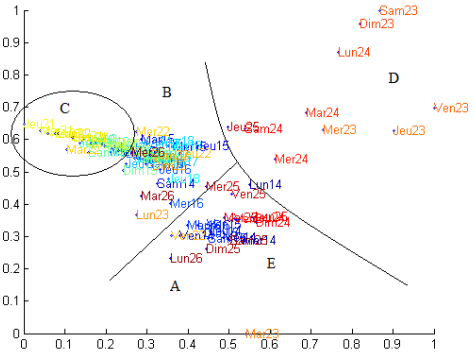
(a) Évolution temporelle de la température du capteur PT325



(b) Projection de la distance ED



(c) Projection de la distance CID



(d) Projection de la distance AFBDTW

FIGURE 5 – Comparaison des différentes métriques de mesure de similarité

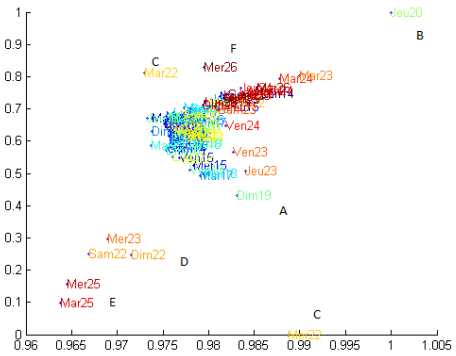


FIGURE 7 – Projection des distances multidimensionnelles *Eros* entre les données présentées en Fig. 6(a)

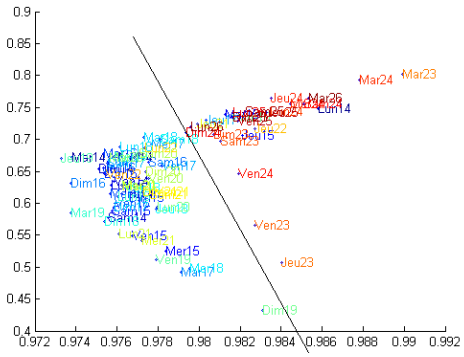
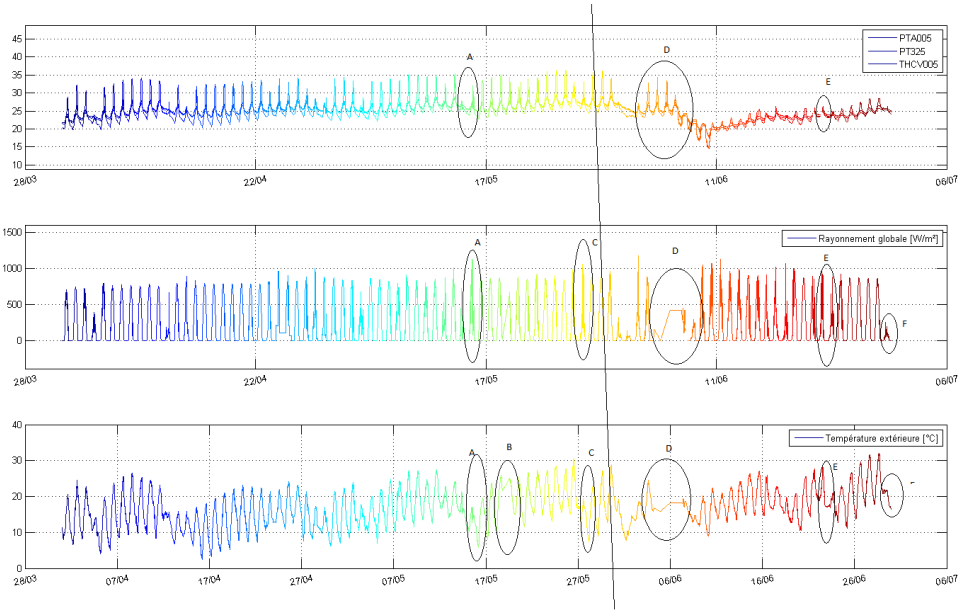
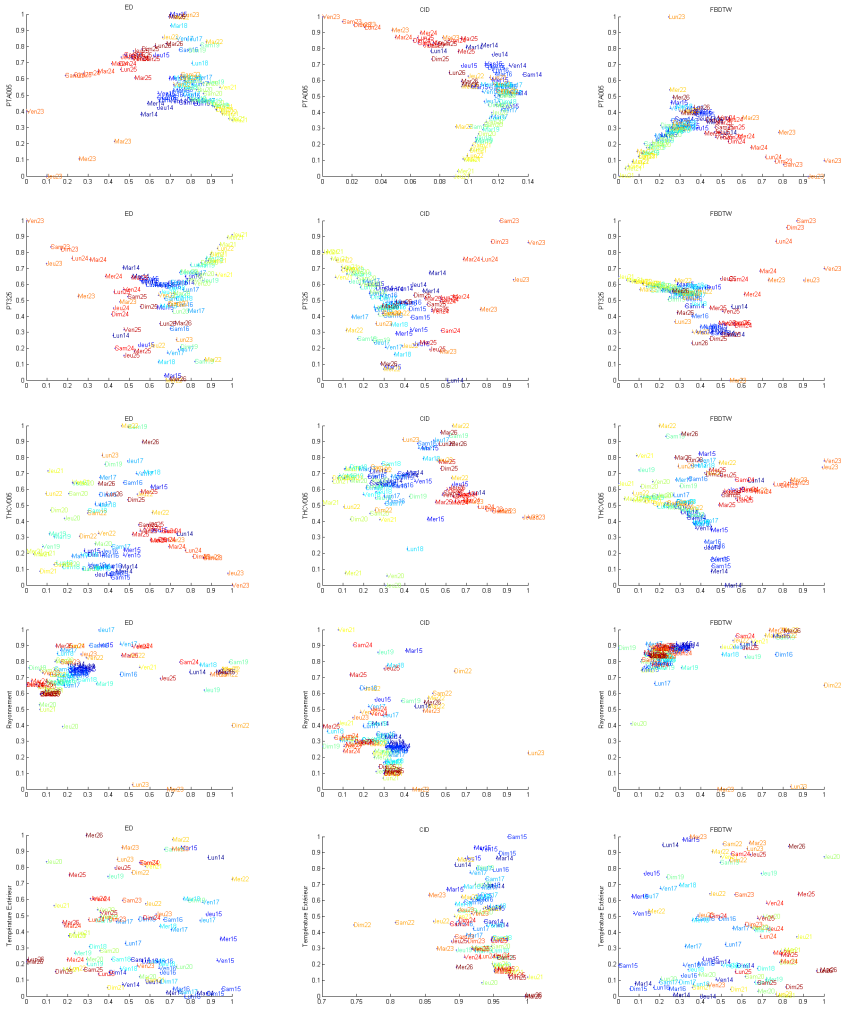


FIGURE 8 – Projection des distances *Eros* sans les jours atypiques

démarche monodimensionnelle, cet ensemble de capteurs engendrera une multitude de projections visuelles à analyser par les experts selon la métrique et le capteur (cf. figure 6(b)). Chaque visualisation isole un comportement différent et d'une façon indépendante de la zone thermique et des mesures ambiantes. Les différentes projections des mesures de rayonnement présentent une concentration localisée des journées contrairement à la température extérieure qui elle se retrouve bien répartie sur le plan de la projection. Or une analyse multidimensionnelle regroupe toutes les informations dans une seule projection. Elle permet entre autres de réduire la dimension des données, d'optimiser l'étude du bâtiment et de faciliter l'identification des anomalies. Les mesures de similarité multidimensionnelles projetées en figure 7 isolent les journées atypiques,



(a) Données MDS cuisine rayonnement



(b) Projection des distances ED, CID et AFBDTW appliquées sur un ensemble de cinq signaux différents

FIGURE 6 – Mesure de similarité monodimensionnelle

les "outliers", présentant des défauts d'enregistrement ou des variations différentes du reste des jours. Ces différents jours sont mis en évidence dans les zones temporelles allant de A à F dans la Fig. 6(a). Une visualisation détaillée des données temporelles de la zone C (Mer22) confirme l'existence de défauts de mesures au niveau de la température extérieure et au niveau du taux de rayonnement global ainsi qu'un fonctionnement normal dans le bâtiment. Ces projections seront à l'origine d'un outil automatique de supervision et d'analyser de données énergétiques. Contrairement aux différentes mesures de similarité de la Fig. 6(b), la projection de la distance multidimensionnelle a permis de réduire d'une façon considérable le volume des données et d'isoler des journées atypiques. Après l'élimination de ces jours, la projection des distances regroupe l'ensemble des journées en deux classes celles des mois d'avril mai à gauche et le mois de juin à droite dans la figure 8. Ces deux groupes présentent des fréquences de variations assez distinctes.

5 Conclusion

Dans ce papier, nous avons proposé plusieurs métriques pour l'aide à l'analyse des données énergétiques du bâtiment. Ces données sont de très grandes dimensions et de natures hétérogènes. Pour faciliter la compréhension et l'utilité de ces métriques, nous avons comparé différentes techniques de mesures de similarité entre des séries temporelles monodimensionnelles. Chaque métrique permet d'isoler un phénomène physique différent. Cependant pour faciliter le déploiement des méthodes automatiques d'analyse de zones thermiques, d'optimisation de l'énergie et plus généralement la prédiction de fonctionnement, l'utilisation d'une métrique multidimensionnelle permet de réduire considérablement la taille des données à traiter et la complexité de l'analyse visuelle. Le choix diversifié des métriques permettra ainsi d'améliorer la qualité des classifieurs fortement impactés par la mesure de similarité.

Pour répondre à la problématique initiale d'optimisation énergétique du bâtiment, la projection de ces métriques avec l'algorithme Isomap peut être couplée avec un outil interactif d'analyse visuelle ou un outil de modélisation thermique pour la maîtrise des réponses énergétiques du bâtiment.

Les outils développés sont également appliqués sur les données issues du programme HOMES [1]. Les résultats préliminaires mettent en avant leur capacité à mettre en évidence les dysfonctionnements liés aux usages de bâtiments occupés.

Références

[1] ADEME (2009), Les chiffres clés du bâtiment, Energie, Environnement, *ADEME Edition,* 2009.

[2] La Tosa V., Marié S., Bernier F. and Piette D., Pervasive Energy Measurements For Buildings Monitoring, in proceedings of the 2nd Workshop on eeBuildings Data Models, CIB W078 - W102, .2011

[3] http ://www.solar-event.com/france/demo/page/index.html

[4] C. Gouy-Pailler, H. Najmeddine, A. Mouraud, F. Suard, C. Spitz, A. Jay and P. Maréchal, Exploring INCAS : Multivariate Data Mining techniques for sensor selection in low-energy consumption and passive buildings, *Proceedings of the CIB W78-W102 2011 : International Conference - Sophia Antipolis*, 2011.

[5] R. Agrawal, C. Faloutos and A. Swami, Efficient Similarity Search In Sequence Database, *Research Report, IBM Almaden Research Center, San Jose, California,* 1993.

[6] Gustavo E.A.P.A. Batista, Xiaoyue Wang, Eamonn J. Keogh, A Complexity-Invariant Distance Measure for Time Series, *SDM*, 2011.

[7] T.Warren Liao, Clustering of time series data - a survey, *The Journal of the Pattern Recognition Society,* Vol. 38, pp. 1857 - 1874, 2005.

[8] D. Sankoff and J. Kruskal, Time warps, string edits, and macromolecules : the theory and practice of sequence comparison, *Addison-Wesley*,1983.

[9] Eamonn J. Keogh and Michael J. Pazzani, Derivative Dynamic TimeWarping, *In First SIAM International Conference on Data Mining SMD*, 2001.

[10] Ying Xie and Bryan Wiltgen, Adaptive Feature Based Dynamic Time Warping, *IJCSNS International Journal of Computer Science and Network Security*, Vol. 10, No.1, 2010.

[11] Kiyoung Yang and Cyrus Shahabi, A multilevel distance-based index structure for multivariate time series, *12th International Symposium on Temporal Representation and Reasoning,* 2005.

[12] N.E Heckman and R.H. Zamar, Comparing the shapes of regression functions, *Biometrika 22, pp. 135-144*, 2000.

[13] J. B. Tenenbaum, V. de Silva and J. C. Langford, A Global Geometric Framework for Nonlinear Dimensionality Reduction, *Science*, 290 (5500) : 2319-2323, 22 December 2000.

[14] C. Wang and X. Sean Wang, Supporting Content-based Searches on Time Series via Approximation, *International Conference on Scientific and Statistical Database Management,* Vol. 69-81, 2000.

[15] Eamonn J. Keogh and M. Pazzani An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback, *In proceedings of the 4th International Conference on Knowledge Discovery and Data Mining,* 1998.

[16] C. A. Ratanamahatana, and Eamonn J. Keogh, Making Time-series Classification More Accurate Using Learned Constraints, *In proceedings of SIAM International Conference on Data Mining (SDM'04), Lake Buena Vista, Florida*, pp.11-22, 2004

[17] S. Park, W. W. Chu, J. Yoon and C. Hsu, Efficient Searches for Similar Subsequences of Different Lengths in Sequence Databases, *16th International Conference on Data Engineering*, pp. 23-32, 2000.

[18] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, *Wiley Interscience, second edition,* 2001.

1. http ://www.homesprogramme.com/