Classification des journées en fonction des radiations solaires sur l'île de la Réunion

Miloud Bessafi, Francisco de A.T. De Carvahlo, Philippe Charton, Mathieu Delsaut, Thierry Despeyroux, Patrick Jeanty, Jean-Daniel Lan Sun Luk, Yves Lechevallier, Henri Ralambondrainy, Lionel Trovalet

Classification des journées en fonction des radiations solaires sur l'île de la Réunion

Miloud Bessafi [5] & Francisco de A. T. de Carvalho [1] & Philippe Charton [4] & Mathieu Delsaut [5] & Thierry Despeyroux [2] & Patrick Jeanty [5] & Jean Daniel Lan-Sun-Luk [5] & Yves Lechevallier [2] & Henri Ralambondrainy [4] & Lionel Trovalet [5]

[1] *CIn/UFPE, Recife-PE, Brésil*
fatc@cin.ufpe.br
[2] *INRIA, Paris-Rocquencourt - 78153 Le Chesnay cedex, France*
{*Yves.Lechevallier,Thierry.Despeyroux*}*@inria.fr*
[4] *LIM, Université de la Réunion-97490 Sainte-Clotilde, Réunion*
{*ralambon,charton*}*@univ-reunion.fr*
[5] *LE²P, Université de la Réunion-97490 Sainte-Clotilde, Réunion*
{*bessafi,mathieu.delsaut,patrick.jeanty,lanson,lionel.trovalet*}*@univ-reunion.fr*

Résumé. L'objectif de cet article est de montrer les intérêts et les inconvénients de deux approches classificatoires de courbes. La première est basée sur une représentation des courbes sous forme vectorielle, la seconde propose la distance de D'Urso et Vichi qui est basée sur les première et seconde dérivées finies. Cette dernière intègre au mieux les propriétés mathématiques des courbes. Ces deux approches seront appliquées à la classification de sources de production d'énergie de type photovoltaïque.

Mots-clés. classification automatique, analyse de courbes.

Abstract. The objective of this paper is to show interest and disadvantages of two approaches for classifying curves. The first is based on a vector representation of curves, the second offers the D'Urso and Vichi distance incorporating the mathematical properties of curves and based on the first and second finite derivatives. These two approaches will be applied to the classification of sources of solar energy.

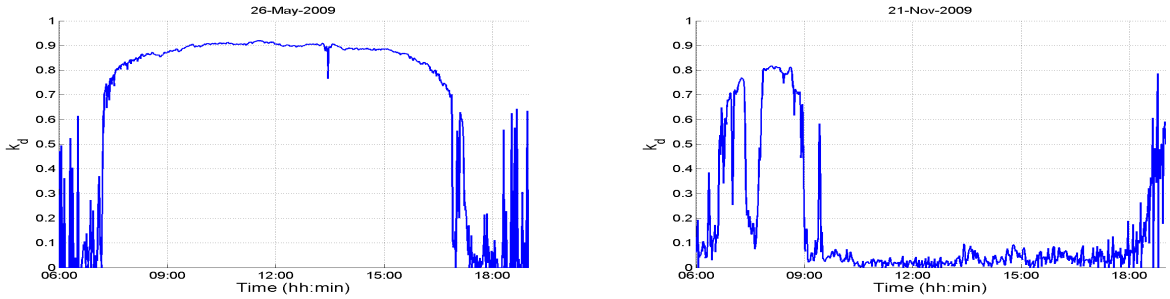
Keywords. clustering, curves analysis.

1 Introduction

L'objectif de ce texte est de comparer deux approches de classification appliquées au partitionnement d'un ensemble de courbes décrivant le rayonnement solaire journalier. Les sources de production d'énergie autonomes intermittentes, de type photovoltaïque, connaissent un développement important à la Réunion. Le laboratoire LIM et ses partenaires se proposent d'améliorer la capacité à prédire la production d'énergie d'une installation photovoltaïque grâce à un réseau de capteurs intelligents. La première étape de cette

démarche consiste au prétraitement de l'information par la classification, (Soubdhan et al.,2009). L'indice de fraction directe noté k_b choisi pour représenter le rayonnement solaire journalier a pour expression : $k_b = \frac{\text{Flux solaire direct}}{\text{Flux solaire global}}$.

Lorsque cet indice est proche de 1, le flux direct est proche du flux global et on est en présence d'une journée ensoleillée; inversement, lorsque l'indice est proche de 0, la journée est nuageuse (Figure 1). L'objectif est de trouver des profils de journées types en tenant compte de leurs degrés d'ensoleillement à partir de k_b. L'échantillon analysé est constitué de 956 journées, du 2008-12-21 au 2012-03-21, sur lesquelles ont été mesurées 9 valeurs horaires de k_b entre 8h et 17h. Ces plages horaires ont été choisies pour éliminer les faibles valeurs de k_b en début et fin de journée.



Une journée ensoleillée de 6h à 19h

Une journée nuageuse de 6h à 19h

Figure 1: Exemples de courbes de l'indice de fraction directe solaire k_b

2 Première approche

Dans cette section, nous présentons la première stratégie de classification adoptée. Soient $I = \{1, \ldots, n\}, T = \{1, \ldots, p\}$, et un ensemble de courbes $E = \{e_i | i \in I\}$ dont les valeurs sont représentées par une matrice $X = (x_i^t)_{i \in I, t \in T}$ où x_i^t est la valeur de la courbe e_i à l'instant t, le poids $p_i = \frac{1}{n}$ est associé à e_i. L'ensemble des données est considéré comme un nuage de points pondérés $\mathbf{N}(I) = \{(\mathbf{x}_i, p_i)\}_{i \in I} \subset \mathbf{R}^p$. La première démarche adoptée pour classer l'ensemble E consiste en la combinaison des trois méthodes éprouvées d'Analyse des Données suivantes:

1. *L'Analyse en Composantes Principales* est utilisée comme une méthode de pré-traitement des données pour éliminer le bruit, réduire la dimension des données en permettant de sélectionner un ensemble réduit de facteurs centrés et non corrélés. Une classification calculée sur un ensemble pertinent de composantes principales est plus stable que celle qui aurait été déterminée sur les données initiales.
2. *La Classification Hiérarchique Ascendante (CAH) de Ward.* Afin de trouver le nombre

optimal de classes pour la partition, la CAH de Ward est appliquée sur ces composantes principales. Cette méthode organise les objets en une suite de partitions emboîtées formant une hiérarchie. A chaque étape, elle réunit deux classes qui minimisent la réduction de l'inertie inter-classes. Un bon nombre de classes peut être déterminé par l'analyse de la décroissance de l'inertie inter-classes des partitions de l'arbre hiérarchique.
3. *La méthode de partitionnement k-means.* Une fois le bon nombre de classes trouvé, la qualité de la partition obtenue par la CAH est améliorée en appliquant la méthode k-means qui maximise l'inertie inter-classes.
Nous avons utilisé la librairie *FactoMineR* (Lê, 2008) qui implémente cette stratégie dans le logiciel *R*.

3 Seconde approche

La seconde approche consiste à utiliser directement une méthode de partitionnement basée sur un ensemble de tableaux de dissimilarités pondérés. Une première approche (E. Diday, G. Govaert, 1977) a été la Classification Automatique avec Distances Adaptatives. La plus récente est la méthode CARD (Clustering and Aggregation of Relational Data) de (H. Frigui et al., 2007) qui introduit une estimation des pondérations pour chaque matrice des dissimilarités. Nous proposons une nouvelle méthode issue de (De Carvalho et al., 2012) pour classer notre ensemble de courbes. Cette méthode permet de partitionner un ensemble d'objets en fonction d'une description basée sur plusieurs matrices de dissimilarités. Chaque classe est représentée par un élément de l'ensemble des objets à classer.

Soient p matrices de dissimilarités $(\mathbf{D}_1, \ldots, \mathbf{D}_j, \ldots, \mathbf{D}_p)$ où $\mathbf{D}_j[i,l] = d_j(e_i, e_l)$ donne la dissimilarité entre les objets e_i et e_l dans la matrice $\mathbf{D}_j$.

Cet algorithme est une version des Nuées Dynamiques qui cherche une partition $P = (C_1, \ldots, C_K)$ de E en K classes, un prototype $g_k \in E$ pour chaque classe C_k et une matrice $\boldsymbol{\lambda}$ des pondérations pour chaque classe et pour chaque matrice de dissimilarités de telle façon que le critère d'adéquation J soit localement optimisé.

$$J(P, \boldsymbol{\lambda}, \mathbf{g}) = \sum_{k=1}^{K} \sum_{e_i \in C_k} d_{\boldsymbol{\lambda}_k}(e_i, g_k) = \sum_{k=1}^{K} \sum_{e_i \in C_k} \sum_{j=1}^{p} \lambda_{kj} d_j(e_i, g_k) \tag{1}$$

dans lequel $d_{\boldsymbol{\lambda_k}}(e_i, g_k)$ est la dissimilarité entre un objet $e_i \in C_k$ et le prototype $g_k \in E$ de la classe paramétrisé par le vecteur de pondération $\boldsymbol{\lambda_k} = (\lambda_{k1}, \ldots, \lambda_{kj}, \ldots, \lambda_{kp})$. Notre algorithme alterne les trois étapes suivantes:

- ***Étape 1: Recherche des meilleurs vecteurs prototypes***

  Dans cette étape, la partition P et la matrice de pondération $\boldsymbol{\lambda}$, sont fixées. Pour chaque classe C_k on recherche le prototype $g_k = e_l \in E$ qui minimise le critère J. Ce prototype est obtenu par: $l = \arg\min_{1 \leq h \leq n} \sum_{e_i \in C_k} \sum_{j=1}^{p} \lambda_h^j \, d_j(e_i, e_h)$.

- ***Étape 2: Calcul de la meilleure matrice de pondération***

  Dans cette étape, la partition P et le vecteur de prototypes $\mathbf{g}$ sont fixés. La pondération λ_{kj} minimisant le critère J avec les contraintes $\lambda_{kj} > 0$ et $\prod_{j=1}^{p} \lambda_{kj} = 1$, est calculée par:

$$\lambda_{kj} = \frac{\left\{\prod_{h=1}^{p}\left[\sum_{e_i \in C_k} d_h(e_i, g_k)\right]\right\}^{\frac{1}{p}}}{\left[\sum_{e_i \in C_k} d_j(e_i, g_k)\right]}$$

- ***Étape 3: Construction de la meilleure partition***

  Dans cette étape, le vecteur de prototypes $\mathbf{g}$ et la matrice de pondération $\boldsymbol{\lambda}$ sont fixés. La classe C_k est construite en utilisant la règle d'affectation suivante:
  $C_k = \{e_i \in E : d_{\boldsymbol{\lambda_k}}(e_i, g_k) < d_{\boldsymbol{\lambda_k}}(e_i, g_h) \forall h \neq k \}$

Il est facile de montrer que chacune de ces trois étapes fait décroître le critère J. L'algorithme démarre avec une partition initiale et alterne ces trois étapes jusqu'à convergence. Cette convergence est atteinte quand la valeur du critère $J(P, \boldsymbol{\lambda_k}, \mathbf{g})$ est stationnaire.

4 Interprétation des classes

L'application de la première stratégie de classification a déterminé une partition a 5 classes de l'indice de fraction directe. Elle sera considérée comme la partition *a priori*. Ces classes résument les régimes d'ensoleillement entre 8h et 17h à l'île de la Réunion, indépendamment de la saison. La partie gauche de la figure 2 donne la tendance moyenne de cet indice pour chaque classe. La partie droite représente l'indice $t_0 = \frac{moy_l(e^t) - moy(e^t)}{\sigma(e^t)}$. Plus la valeur absolue est grande plus la variable joue un rôle discriminant pour cette classe. Le signe de cette valeur positionne la classe par rapport à la valeur centrale.

Cette figure fait apparaître deux types de courbes. La première concerne celles qui commencent avec un ciel relativement bien dégagé faisant prévaloir des valeurs de k_b élevées, ce sont les courbes des classes 2,4 et 5. La seconde concerne les courbes des classes 1 et 3 ayant des faibles valeurs de k_b ce qui laisse entrevoir des débuts de journée avec ciel très couverts.

Classe 1 (C_1): : journées nuageuses. Effectif : 146, 15%. Cette classe correspond à un niveau d'ensoleillement très faible toute la journée. Le faible niveau de la valeur moyenne de K_b indique une couverture nuageuse importante. Cette classe montre des phénomènes locaux dominants parmi lesquels on peut citer les faibles alizés en été austral, les flux d'humidité importants et les brises de terre induites par des contrastes thermiques importants notamment en été.

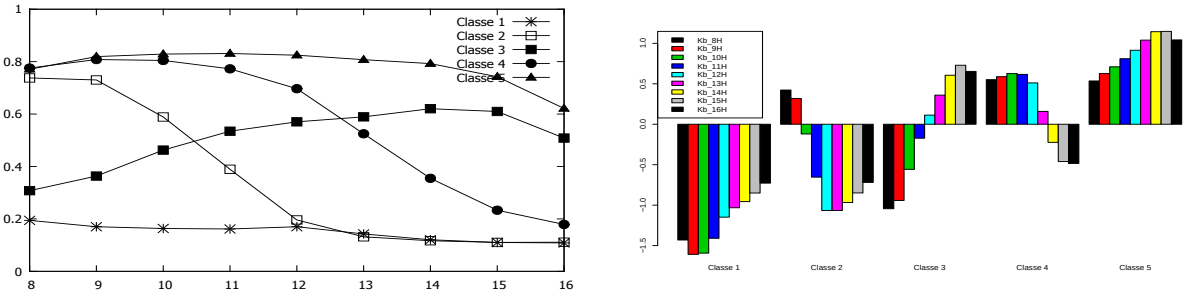


Figure 2: Tendances du K_b et du t_0 pour chaque classe de la partition a priori

Classe 2 (C_2) : journées intermédiaires mauvaises. Effectif : 189, 19.7%. La classe 2 présente une matinée ensoleillée jusqu'en milieu de matinée vers 9h-9h30 et un après-midi très nuageux. C'est le régime de temps classique de l'été austral.

Classe 3 (C_3) : journées perturbées. Effectif : 131, 13.7%. La classe 3 correspond à une journée variable avec une amélioration du temps en fin de matinée et une couverture nuageuse modérée dans l'après-midi.

Classe 4 (C_4) : journées intermédiaires bonnes. Effectif : 232, 24%. Le comportement de la classe 4 est similaire à celui de la classe 2, cependant le régime ensoleillé est plus marqué jusqu'au début de l'après-midi.

Class 5 (C_5) : journées ciel clair. Effectif : 258, 26.9%. La classe 5, correspond à un régime de beau temps sur toute la journée avec un rayonnement direct qui prédomine.

Dans la seconde approche la partition DD (Figure 3) est obtenue en utilisant la distance de d'Urso et Vichi (D'Urso et Vichi, 1998) qui est basée sur trois dissimilarités, la première compare les valeurs des plages horaires de cette courbe, la seconde mesure la vélocité (la dérivée discrète entre deux positions) et la troisième mesure l'accélération (dérivée seconde discrète). Elle est bien adaptée pour mesurer la proximité entre deux courbes.

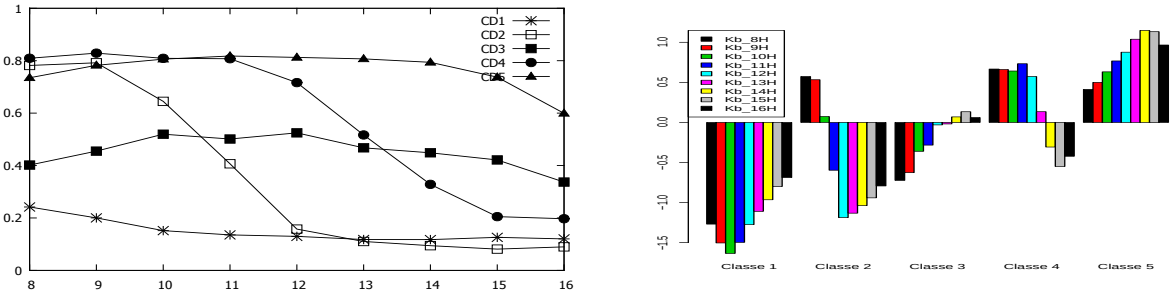


Figure 3: Tendances du K_b et du t_0 pour chaque classe de la partition DD

La relation entre les classes de la partition *a priori* (Figure 2) et les classes de la partition DD (Figure 3) est effectuée par minimisation de l'erreur de classification (MAP). La partition DD est très proche de la partition a priori car l'erreur globale de classification (OERC) est de 82,95%. Correspondant à l'erreur de classification, la ligne "Rappel" de la table 1 donne un rappel assez semblable pour les classes extrêmes (84,04% pour la classe 1 et 97.29% pour la classe 5). Par contre les trois classes intermédiaires ont un rappel proche de 75%.

PP	C_1	C_2	C_3	C_4	C_5	Sum	%
CD_1	**130**	17	7	0	0	154	16,11%
CD_2	0	**137**	0	1	0	138	14,44%
CD_3	16	32	**95**	35	3	181	18,93%
CD_4	0	3	2	**180**	4	189	19,77%
CD_5	0	0	27	16	**251**	294	30,75%
Sum	146	189	131	232	258	956	
%	15,27%	19,77%	13,70%	24,27%	26,99%		
Rappel	**89,04%**	72,49%	72,52%	77,59%	**97,29%**	**82,95%**	

Table 1: Tableau de confusion entre la partition a priori et la partition DD

Conclusion

L'utilisation d'un ensemble de matrices de distances permet d'intégrer facilement la structure complexe des courbes et de donner un rôle collaboratif à chaque matrice de dissimilarité. Les prototypes sont des courbes issues de notre population, ce qui donne une interprétation simple d'une classe. Les pondérations sont optimisées localement par classe et par matrice de dissimilarités.

Bibliographie

[1] F. A. T. De Carvalho, Y. Lechevallier, and F. M. De Melo, Partitioning hard clustering algorithms based on multiple dissimilarity matrices. *Pattern Recognition* 45, 447-464, 2012

[2] E. Diday, and G. Govaert, Classification automatique avec distances adaptatives. *R.A.I.R.O. Informatique Computer Science* 11(4), 329-349, 1977

[3] H. Frigui, C. Hwang, F. C.-H. Rhee, Clustering and aggregation of relational data with applications to image database categorization. *Pattern Recognition* 40, 3053-3068, 2007

[4] S. Lê and J. Josse and F. Husson, FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software* 25(1), 1-18, 2008

[5] T. Soubdhan and R. Emilion and R. Calif, Classification of daily solar radiation distributions using a mixture of Dirichlet distributions. *Solar Energy* 83(7), 1056-1063, 2009

[6] P. D'Urso and M. Vichi, Dissimilarities between trajectories of a three-way longitudinal data set, In A. Rizzi, M. Vichi, H.-H. Bock, *Advances in Data Science and Classification*, Springer, Berlin, 585-592, 1998