Classification et prédiction du flux solaire

Henri Ralambondrainy*, Yves Lechevallier**, Jean Daniel Lan-Sun-Luk***, Jean-Pierre Chabriat***

*LIM, Université de la Réunion-97490 Sainte-Clotilde, Réunion
ralambon@univ-reunion.fr,
**INRIA, Paris-Rocquencourt - 78153 Le Chesnay cedex, France
Yves.Lechevallier@inria.fr
***LE^{2}P, Université de la Réunion-97490 Sainte-Clotilde, Réunion
{lanson,doyensc}@univ-reunion.fr

Résumé. La prédiction du rayonnement solaire horaire dans une journée est un enjeu primordial pour la production d'énergie de type photovoltaïque. Nous présentons deux stratégies de classification des jours selon leurs rayonnements solaires puis une méthode de prédiction du flux solaire cohérente avec la classification.

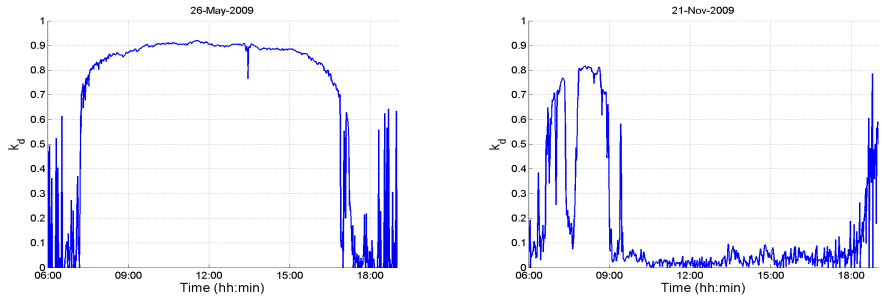
1 Introduction

Les sources de production d'énergie autonomes intermittentes, de type photovoltaïque, connaissent un développement important dans les îles subtropicales. Un projet a été mis en place pour améliorer la capacité à prédire la production d'énergie d'une installation photovoltaïque grâce à un réseau de capteurs intelligents. Les données disponibles concernent 956 journées, du 2008-12-21 au 2012-03-21, sur lesquelles ont été mesurés les cumuls horaires du rayonnement solaire journalier de 9H jusqu'à 17h. Nous présentons deux stratégies (Bessafi et al., 2013) de classification des jours selon leurs rayonnements solaires puis une méthode de prédiction du flux solaire basée sur les résultats des classifications précédentes

2 Classification

Le rayonnement solaire peut-être décomposé en trois flux : le flux global F_{Global}, diffus F_{Diffus} et direct $F_{Direct} = F_{Global} - F_{Diffus}$. Nous définissons l'indice de fraction directe noté $k_b = F_{Direct}/F_{Global}$ pour représenter le rayonnement solaire journalier. 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). On note I l'ensemble de n journées, T l'ensemble de p heures et K le nombre de classes. Dans la suite, les indices i, t décriront respectivement I, T et $k = 1 \dots K$. Dans la première approche, une journée d_i par le vecteur des indices k_b horaires $x_i = (k_b(i,t))_t$.

La première démarche pour classer l'ensemble des journées combine trois méthodes éprouvées d'analyse des données :



Une journée ensoleillée de 6h à 19h Une journée nuageuse de 6h à 19h

FIG. 1 – *Exemples de courbes de l'indice de fraction directe solaire* k_b

- *l'Analyse en Composantes Principales* pour réduire la dimension des données. Afin de trouver un nombre optimal de classes pour la partition,
- *la Classification Hiérarchique Ascendante de Ward (CAH)* est appliquée sur un ensemble pertinent de composantes principales.
- la qualité de la partition obtenue par la CAH est ensuite améliorée en appliquant la méthode *k-means*. La librairie *FactoMineR* (Lê et al., 2008) qui implémente cette stratégie dans le logiciel *R* a été utilisée.

Dans la seconde approche, une journée d_i est caractérisée par trois composantes $d_i = (x_i, v_i, a_i)$ où $x_i = (k_b(i,t))_t$ est *la position*, $v_i = (\frac{\Delta x(i,t)}{\Delta t})_t$ la *vélocité* et $a_i = (\frac{\Delta^2 x(i,t)}{\Delta t^2})_t$ *l'accélération*. Pour classifier les journées, un indice global de dissimilarité est défini sur les paires des journées d_i et d_l par $D^2(d_i, d_l) = \lambda_1.D_1^2(x_i, x_l) + \lambda_2.D_2^2(v_i, v_l) + \lambda_3.D_3^2(a_i, a_l)$ où $\lambda_1, \lambda_2, \lambda_3$ sont des pondérations et D_1, D_2, D_3 les indices de dissimilarité définis respectivement sur ces trois vecteurs. Si D_1, D_2, D_3 sont les distances euclidiennes usuelles, cet indice est appelé l'indice de dissimilarité d'Urso and Vichi (D'Urso et Vichi, 1998). La difficulté dans l'utilisation de cet indice réside dans le choix des pondérations des dissimilarités.

La plus récente méthode pour la détermination de poids optimaux est la méthode CARD (Clustering and Aggregation of Relational Data) de (Frigui et al., 2007) qui introduit une estimation des pondérations pour chaque matrice des dissimilarités. Nous proposons une nouvelle méthode (De Carvalho et al., 2012) qui détermine simultanément un ensemble de pondérations optimales et une classification des objets décrits par plusieurs matrices de dissimilarités.

Soient $P = (C_1, \ldots, C_K)$ une partition de E, un ensemble de matrices de dissimilarités D_t définies sur E et une matrice $\boldsymbol{\lambda} = (\lambda_{kt})_{k,t}$ où λ_{kt} est la pondération associée à la dissimilarité D_t et à la classe C_k. Chaque classe C_k est représentée par un prototype $g_k \in E$ et $G = (g_1, \ldots, g_K)$ le vecteur des prototypes.

Le problème de classification s'énonce comme la recherche du triplet optimal $(P^*, G^*, \boldsymbol{\lambda}^*)$ solution de : $(P^*, G^*, \boldsymbol{\lambda}^*) = arg\ min_{P,G,\boldsymbol{\lambda}} W(P, G, \boldsymbol{\lambda})$ où

$$W(P, G, \boldsymbol{\lambda}) = \Sigma_{k=1}^{K} \Sigma_{d_i \in C_k} \Sigma_{t \in T} \lambda_{kt} D_t^2(d_i, g_k).$$

L'algorithme proposé comporte trois étapes.

- *Étape 1 : construction de la meilleure partition en K Classes*

- *Étape 2 : calcul de la meilleure matrice de pondération*
- *Étape 3 : recherche le meilleur vecteur de K prototypes*

L'algorithme démarre avec un vecteur de prototypes tiré au hasard et toutes les pondérations égales à 1 et alterne ces trois étapes jusqu'à la convergence.
L'application de la première stratégie classique de classification sur la composante *position* a déterminé une partition P_1 a 5 classes des journées. Pour étudier l'influence des composants *vélocité* et *accélération*, nous appliquons la seconde stratégie qui calcule les pondérations $\boldsymbol{\lambda_{t,k}}$ pour chaque classe et pour chaque dissimilarité D_t et détermine la partition P_2. Les courbes des moyennes des classes des partitions P_1 et P_2 sont similaires (figure 2). La similitude des deux partitions est confirmée par l'analyse du tableau 1 de confusion entre P_1 et P_2. L'erreur globale est de 17,05, les classes extrêmes 1 et 5 ont un bon score de rappel de 89,04% et 97.29%, les classes intermédiaires de 75%.

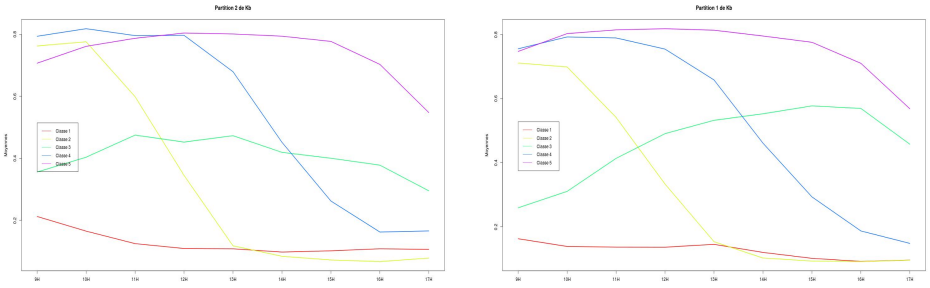


FIG. 2 – *Moyennes de k_b pour les classes des partitions P_1 et P_2.*

P_2 \ P_1	1	2	3	4	5	Total	%
1	**130**	17	7	0	0	154	16,11%
2	0	**137**	0	1	0	138	14,44%
3	16	32	**95**	35	3	181	18,93%
4	0	3	2	**180**	4	189	19,77%
5	0	0	27	16	**251**	294	30,75%
Total	146	189	131	232	258	956	
%	15,27%	19,77%	13,70%	24,27%	26,99%	**17,05%**	
Rappel	**89,04%**	72,49%	72,52%	77,59%	**97,29%**	**82,95%**	

TAB. 1 – *Tableau de confusion entre les partitions P_1 et P_2.*

- **Classe 1 : journées nuageuses.** Effectif : 146, 15,27%. 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.

- **Classe 2 : journées intermédiaires mauvaises.** Effectif : 189, 19.77%. La classe 2 présente une matinée ensoleillée jusqu'en milieu de matinée vers 10h-10h30 et un après-midi très nuageux. C'est le régime de temps classique de l'été austral.
- **Classe 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 : journées intermédiaires bonnes.** Effectif : 232, 24,27%. 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.
- **Classe 5 : journées ciel clair.** Effectif : 258, 26.99%. La classe 5, correspond à un régime de beau temps sur toute la journée avec un rayonnement direct qui prédomine.

3 Prédiction

L'objectif du projet est de proposer des outils de prédiction, à travers un site web, du rayonnement solaire horaire dans une journée. L'intérêt de ces prédictions est de permettre d'anticiper un pic ou une chute de la production d'électricité pour l'heure à venir. Dans la suite, F désignera un des indicateurs de flux $F_{Global}, F_{Diffus}, K_b$. Notons $F(i,t)$ le flux à prévoir à l'heure t d'une journée d_i et $\widehat{F}(i,t)$ la valeur estimée. Le problème de prédiction se formule comme la recherche d'une fonction f telle que $\widehat{F}(i,t) = f(F(i,1:(t-1)))$. La classification précédente a mis en évidence l'existence de plusieurs régimes de flux solaire journalier (figure 2). Elle suggère de rechercher des modèles de prévision horaire par classe qui seraient plus appropriés qu'un modèle unique.

Si $P = \{C_k\}_k$ est une partition des jours, le modèle de prévision locale s'écrit : $\widehat{F}(i,t) = \Sigma_k f_k(F(i,1:t-1))\mathbf{1}_{C_k}(d_i)$. La mise en oeuvre de cette approche nécessite le choix d'une partition P, d'une fonction d'affectation $\mathbf{1}_{C_k}$ d'une observation à une classe et d'un modèle de prévision par classe f_k. *Cart-Regression* (Breiman et al., 1984) est un exemple de méthode qui adopte cette approche locale. Son modèle de prévision est $\widehat{F}(i,t) = \Sigma_k \text{Moy}_k F(t)\mathbf{1}_{C_k}(d_i)$. La moyenne du flux $F(t)$ dans la classe C_k est l'estimation de la valeur du flux pour les journées de cette classe. Des partitions homogènes relatives à *la variable à prédire* $F(t)$ sont déterminées de manière récursive et des arbres de décision calculés sur les variables prédictives permettent d'affecter une journée aux classes de ces partitions.

La méthode de prévision globale (*Regr-Globale*) que nous avons utilisée s'appuie sur un modèle linéaire simple : $\widehat{F}(i,t) = a(t) + b(t) \times F(i,t-1)$. Plusieurs types de coefficients ont été essayés, comme $a(t) = 0, b(t) = \frac{\text{Moy}F(t)}{\text{Moy}F(t-1)}$ ainsi que d'autres statistiques la médiane, le troisième quartile. La régression linéaire simple a été retenue car elle a donné le meilleur score de prévision sur les ensembles de test. Ce modèle de régression linéaire simple a été aussi choisi pour l'approche locale. Ce choix a été motivé par la propriété physique de la persistance du flux horaire, la contrainte du projet d'avoir un système de prédiction efficace en ligne et les études préalables sur la sélection de variables discriminantes pour la régression. Le modèle de prévision de la régression locale (*Regr-Locale*) proposée s'écrit $\widehat{F}(i,t) = \Sigma_k(a(k,t) + b(k,t) \times F(i,t-1))\mathbf{1}_{C_k}(d_i)$ où $a(t,k), b(t,k)$ sont les coefficients estimés pour la classe C_k d'une partition P de l'échantillon d'apprentissage des jours.

Pour cette méthode, une partition *unique* des jours caractérisés par ses flux horaires est déterminée par la première méthodologie de classification précédente. La qualité d'une méthode de prédiction est mesurée par le rapport des normes $R = \| \widehat{F} \| / \| F \| = \sqrt{\frac{\Sigma_i \Sigma_t \widehat{F}(i,t)^2}{\Sigma_i \Sigma_t F(i,t)^2}}$ et l'écart-type de l'écart quadratique moyen $\sigma_{EQM} = \sqrt{\frac{1}{n \times p} \Sigma_i \Sigma_t (F(i,t) - \widehat{F}(i,t))^2}$.

(a) Valeurs moyennes de σ_{EQM} sur les ensembles d'apprentissage.

Apprentissage	10H-17H		
(Effectif : 75% , Essai : 10)	F_{Global}	F_{Diffus}	K_b
Partition : *nombre de classes*	σ_{EQM}	σ_{EQM}	σ_{EQM}
1	134.58	76.85	0.178
3	128.02	72.22	0.167
5	124.37	69.58	0.151
10	117.97	66.43	0.140
15	102.71	58.51	0.132

(b) Valeurs moyennes de σ_{EQM} et R des flux sur les ensembles de test.

Validation croisée	10H-17H		
(Effectif : 25% , Essai : 10)	F_{Global}	F_{Diffus}	K_b
Méthode		σ_{EQM}	
Cart-Regression	139.93	79.99	0.179
Regr-Globale	135.0	76.58	0.177
Regr-Locale	134.7	76.28	0.177
Classe 1	127.07	66.03	0.157
Classe 2	121.21	78.95	0.199
Classe 3	128.95	78.69	0.166
Classe 4	137.59	80.11	0.173
Classe 5	145.09	83.06	0.172
	$R = \| \widehat{F} \| / \| F \|$		
Cart-Regression	0.973	0.95	0.95
Regr-Globale	0.97	0.94	0.947
Regr-Locale	0.97	0.94	0.950

TAB. 2 – *Résultats comparatifs.*

Le tableau 2(a) donne les valeurs de σ_{EQM} des flux relatifs à des partitions calculées sur des échantillons d'apprentissage pour des nombres de classes différents. De manière logique, on constate que plus le nombre de classes K augmente plus la qualité de la prédiction s'améliore (l'indice σ_{EQM} décroit). Pour $K = n$, on a $\sigma_{EQM} = 0$ car à un individu correspond une classe, la prédiction est parfaite mais ne présente pas d'intérêt (sur-apprentissage). Sur les ensembles de test, la classe d'appartenance d'une journée est inconnue. A l'heure t, une journée est affectée à la classe la plus proche selon sa distance aux centres de gravité des classes calculée à partir de $F(1 : t-1)$. Le modèle de régression de la classe d'affectation est ensuite choisi pour l'estimation de $F(t)$.

Le tableau 2(b) donne les scores moyens, sur les ensembles de test, des méthodes *Cart-Regression, Regr-Globale* et *Regr-Locale*. On constate que la qualité moyenne de prévision de

Regr-Globale est équivalente à celle de *Regr-Locale* et qu'elle est meilleure que celle de *Cart-Regression*. L'examen des σ_{EQM} par classe montre que cet indice est meilleur pour certaines classes (classes 1,2,3 pour le flux FG par exemple) que le score moyen. Les valeurs du critère R sont de bonne qualité et équivalentes pour les trois méthodes.

4 Conclusion

Nos approches classificatoires nous ont permis d'obtenir une partition pertinente et interprétable d'un point de vue physique des jours selon un indice de flux solaire. Cette typologie nous a conduit à proposer une méthode de régression locale adaptée aux différents régimes du flux solaire. Le bon score de prévision obtenu pour certaines classes encourage la recherche de meilleurs fonctions discriminantes des classes et d'une méthode de "Classification-Régression" où le couplage classe/modèle de régression serait plus fort.

Références

Bessafi, M., F. A. T. D. Carvalho, P. Charton, M. Delsaut, T. Despeyroux, P. Jeanty, J.-D. Lan-Sun-Luk, Y. Lechevallier, H. Ralambondrainy, et L. Trovalet (2013). Classification des journées en fonction des radiations solaires sur l'île de la réunion. Toulouse : Journées Françaises de Statistique.

Breiman, L., J. H. Friedman, R. A. Olshen, et C. J. Stone (1984). *Classification And Regression Trees*. New York: Chapman and Hall.

De Carvalho, F. A. T., Y. Lechevallier, et F. M. D. Melo (2012). Partitioning hard clustering algorithms based on multiple dissimilarity matrices. *Pattern Recognition 45*, 447–464.

D'Urso, P. et M. Vichi (1998). Dissimilarities between trajectories of a three-way longitudinal data set. In A. Rizzi, M. Vichi, et H.-H. Bock (Eds.), *Advances in data science and classification.*, pp. 585–592. Berlin: Springer.

Frigui, H., C. Hwang, et F. Rhee (2007). Clustering and aggregation of relational data with applications to image database categorization. *Pattern Recognition 40*, 3053–3068.

Lê, S., J. Josse, et F. Husson (2008). Factominer: An R package for multivariate analysis. *Journal of Statistical Software 25*, 1–18.

Summary

The prediction of the solar irradiance in a day is a crucial issue for the energy production by intermittent energy sources. We propose two clustering strategies and a forcasting method for solar irradiance.