Identification de modèles de rayonnement solaire en zone tropicale par critères d'information

Laurent Linguet[1], Yannis Pousset[2], Christian Olivier[2]

1. Laboratoire UMR Espace-DEV, Université de la Guyane
IRD, 275 route de Montabo, BP 165,
97323 Cayenne cedex, Guyane française
laurent.linguet@guyane.univ-ag.fr

2. Laboratoire SIC, Université de Poitiers
Département SIC (Signal, Images et Communication)
Institut XLIM, 86962 Futuroscope Cedex, France
yannis.pousset,christian.olivier@univ-poitiers.fr

RÉSUMÉ. L'objet de cet article est d'améliorer la connaissance du rayonnement solaire en zone tropicale à travers l'analyse des données de mesures d'irradiance au sol. Pour cela nous identifions, à l'aide de critères d'information, les distributions de probabilité introduites dans quelques modèles de génération de rayonnement solaire synthétique. Puis, nous validons les résultats à partir de différentes mesures et différents tests entre distributions issues des données réelles et celles synthétisées.

ABSTRACT. The aim of this article is to improve the knowledge of the solar radiation in the tropics through the analysis of irradiance measurements at ground. For this, we identify probability distributions introduced in some synthetic solar radiation models, using information criteria. Validation is conducted through different tests and measures between real data distributions and synthesized data distribution.

MOTS-CLÉS : critères d'information, modèles de rayonnement solaire synthétique, sélection de modèle.

KEYWORDS: criteria, synthetic solar radiation model, model selection.

DOI:10.3166/TS.31.363-381

Extended abstract

The aim of this article is to improve the knowledge of the solar radiation in the tropical areas, through the analysis of irradiance measurements. For that, we dispose of solar radiation time series. They are obtained from measured data at ground or transmitted by satellite. These data or samples are obviously incomplete, no totally representative, and subject to perturbation.

All these constraints lead to introduce statistical models generating the time series of data. Among several models proposed in literature, we choose the TAG model (for Time dependent Autoregressive Gaussian) of Aguiar and Collares-Pereira in 1992, and the model with high temporal resolution given by Polo and al. in 2011. These model generating sequences present random terms with probability density function which are a priori chosen.

The originality of this paper consists in reconsidering the nature of these probability laws. We dispose of a great number of data. So, to select the best probability law we use a statistical tool based on Information Criteria, also referred as penalized log-likelihood criteria. We retain BIC and Φ_β criteria for their strong consistence.

Among retained probability law candidates, the BIC and Φ_β criteria agree to give the same law for the TAG and the Polo models, and this, whatever the different sky conditions. These conclusions differ in part of the literature.

Then we compare the two probability density functions between solar radiation measured and solar radiation generated from previously elected probability laws. An exact concordance is observed for the TAG model and for the Polo model, except for partially cloud cover case. For this case, the very close values of criteria and the similarity of Nakagami and Beta cumulative functions could explain this discordance.

1. Introduction

La connaissance du rayonnement solaire, ou irradiance, à la surface de la terre est d'un grand intérêt dans de nombreux domaines. Les sciences du climat requièrent des données solaires fiables et suffisamment nombreuses pour comprendre le changement climatique. L'agriculture et plus généralement les écosystèmes naturels sont affectés par le rayonnement solaire et sa connaissance est nécessaire pour aider à la compréhension des actuels impacts liés au changement climatique (Stanhill et Cohen, 2001). Sur le plan énergétique, la conception et le dimensionnement de systèmes utilisant l'énergie solaire en entrée, tels que les chauffe-eau solaires, les cellules photovoltaïques ou les concentrateurs solaires thermiques, nécessitent des données solaires afin de simuler et tester l'efficacité énergétique à long terme de ces systèmes (Mellit *et al.,* 2008). En architecture, la simulation des performances énergétiques des immeubles en zone urbaine requiert aussi de disposer de données de rayonnement solaire (données d'entrée), afin de

dimensionner les systèmes de production d'énergie propre complémentaires (thermique solaire, photovoltaïque, etc.) aptes à satisfaire les besoins en chauffage et en énergie électrique tout en optimisant la consommation totale d'énergie des immeubles (Amado et Poggi, 2012). Dans tous ces domaines et dans d'autres, les données de rayonnement solaire représentées par des séries temporelles à moyen et à long terme sont souvent nécessaires.

Les séries temporelles de rayonnement solaire peuvent être obtenues, soit à partir de données mesurées au sol ou à partir de données mesurées par satellite (Marie-Joseph *et al.*, 2013), soit en sélectionnant des périodes de données représentatives des mesures et en calculant une année moyenne de rayonnement appelée année météorologique typique (*Typical Meteorological Year, TMY*) (Bilbao *et al*., 2004).

Cependant, ces méthodes bien que simples, ont des désavantages. Dans le premier cas les séries temporelles sont limitées à la reproduction de données historiques et ne reproduisent pas la totalité de la plage de variabilité des données de rayonnement, en outre elles peuvent parfois être incomplètes. Dans le second cas, il n'y a aucune garantie à ce que la *TMY* élaborée inclue les caractéristiques statistiques à long terme du climat de la localisation choisie, ni reproduise les valeurs extrêmes de rayonnement. Enfin, autre désavantage, il existe de nombreuses régions où la faible densité de stations de mesures au sol et/ou l'absence de données de rayonnement dérivées de mesures satellite ne permettent pas l'utilisation de ces méthodes. Toutes ces contraintes obligent le plus souvent à recourir à des séries temporelles de rayonnement générées synthétiquement à différentes résolutions temporelles (heures, journées, etc.) en fonction des besoins de l'utilisateur.

Durant ces deux dernières décennies, plusieurs méthodes ont été développées pour générer synthétiquement des séries temporelles de rayonnement solaire. Les modèles utilisés doivent tous respecter la condition suivante (Hansen *et al.*, 2010) : produire des séries ayant le même contenu statistique que celui de séquences temporelles observées sur les localisations d'intérêt. De plus, les valeurs de rayonnement simulées doivent être statistiquement consistantes avec celles mesurées. Parmi les méthodes les plus connues on peut citer : une méthode utilisant un modèle *ARMA* (*Auto Regressive Moving Average*) développée initialement par Graham et Hollands (1988) puis améliorée par Aguiar et Collares-Pereira (1992) et utilisée par d'autres auteurs (Muselli, 1998 ; Tiba, 2004) ; celle utilisant un modèle markovien associé à une matrice de transition (*Markov Transition Matrix Model*) développée par Aguiar *et al.* (1988) et reprise par d'autres auteurs en incluant notamment l'utilisation de réseaux de neurones pour configurer la matrice de transition (Poggi, 2000 ; Linares-Rodríguez, 2011) ; la méthode élaborée par Boland (1995) qui associe modèle autorégressif (*AR*) et analyse de Fourier ; la méthode développée par Polo (2011) qui permet de modéliser une série temporelle à partir d'une valeur moyenne à laquelle on rajoute une fluctuation aléatoire dont les caractéristiques (amplitude et fréquence) dépendent des conditions de nébulosité du ciel.

Les trois premières méthodes utilisent, par facilité, une distribution gaussienne pour représenter le terme de l'erreur (bruit) de modélisation. Ce choix est adopté par

la plupart des outils qui permettent de déterminer les paramètres des modèles autorégressifs (*AR* ou *ARMA*) car, dans le domaine de l'analyse des séries temporelles l'hypothèse d'un bruit blanc gaussien est une condition simplificatrice à la détermination du modèle. Cependant dans la réalité, les séries temporelles de rayonnement solaires sont physiquement bornées : elles ne comportent que des valeurs positives et elles ne peuvent dépasser une valeur maximale correspondant au rayonnement solaire extra-terrestre (rayonnement solaire mesuré au-dessus de l'atmosphère). La distribution gaussienne n'est donc peut-être pas celle qui donne la meilleure représentation de l'erreur associée aux modèles comme l'ont montré Remund (2002) et Boland (2008).

La dernière méthode, celle de Polo, utilise une distribution Bêta pour modéliser la fluctuation aléatoire du rayonnement (le terme aléatoire). Cependant, dans la littérature, outre la distribution Bêta, on retrouve aussi d'autres distributions (Boltzman, Gamma, Exponentielle) pour approximer les lois de probabilité des séries temporelles de rayonnement solaire.

L'objet de ce papier consiste à déterminer les lois de probabilité décrivant le mieux possible la distribution statistique de chaque terme aléatoire intervenant dans deux modèles connus de génération de rayonnement solaire synthétique : le modèle d'Aguiar et Pereira (1992) et le modèle de Polo (2011).

Nous introduirons pour déterminer ces distributions une méthode connue de sélection de lois, basée sur les critères d'information (IC pour *Information Criterion*) appelés aussi critères entropiques généralisés, qui seront préférés au simple maximum de vraisemblance (*cf.* le principe de parcimonie). Nous retiendrons les critères BIC (Schwarz, 1978) et Φ_β (El Matouat et Hallin, 1996) pour leur forte consistance (convergence presque sûre). Ils requièrent en général un assez grand nombre de données, ce dont nous disposons, comme il sera indiqué par la suite lors de la phase expérimentale.

L'organisation de ce papier est la suivante : dans la section 2, nous rappelons les concepts statistiques et physiques utilisés par la suite : les critères d'information *IC*, la génération de séries temporelles de rayonnement solaire ; dans la section 3, nous décrivons le processus d'identification entre diverses lois candidates correspondant aux lois d'erreurs introduites dans les modèles de génération ; dans la section 4, nous analysons les performances des modèles synthétiques issus du processus d'identification des erreurs, en les comparant aux données réelles, de façon à valider ou non les conclusions du processus précédent de sélection de lois.

2. Rappel de quelques concepts statistiques et physiques

2.1. Les critères d'information

Les critères d'information (pour un état de l'art, voir Olivier et Alata, 2009) sont des outils fournissant une réponse partielle au problème de parcimonie : étant donnée une suite de réalisations ou données ou observations $x^N = (x_1, \ldots, x_N)$ d'un

processus aléatoire X, et étant donnée une famille Θ de modèles paramétriques choisie à priori, quel est le modèle $\widehat{\boldsymbol{\theta}}$ de Θ qui correspond le mieux au processus X ? En d'autres termes, cela signifie que l'on recherche le nombre et les valeurs des paramètres libres du modèle $\widehat{\boldsymbol{\theta}}$, optimaux au sens des *IC*.

Le concept des *IC* est le suivant : assigner à chaque modèle θ_i de Θ en compétition une pénalité « compensant » la -log-vraisemblance classique $L(\theta_i)$, ce qui reviendra à minimiser l'expression :

$$IC(i) = L(\theta_i) + |\theta_i| C(N), \tag{1}$$

où $C(N)$ est un terme dépendant en général du nombre d'observations N et $|\theta_i|$ est le nombre de paramètres libres du modèle θ_i. Rappelons que le seul critère du maximum de vraisemblance (ici la minimisation de $L(\theta_i)$) est insuffisant lorsque $|\theta_i|$ varie.

L'expression de la pénalité $|\theta_i|C(N)$ est obtenue à partir de la minimisation d'un coût entre modèles, en général de type *f*-divergence, bayésien ou complexité stochastique, et diffère suivant les critères. Le plus connu et le plus ancien est le critère d'Akaike (*AIC*) (1974). Dans notre étude, nous ne retenons que ceux de Schwarz (1978) et de El Matouat et Hallin (1996), notés respectivement *BIC* et Φ_β, en oubliant les autres critères non consistants fortement, c'est-à-dire non convergeant presque sûrement quand N $\rightarrow +\infty$ (voir Olivier et Alata, 2009)). Ainsi par exemple, le critère de Hannan et Quinn (1979), noté Φ, est faiblement consistant (convergence en probabilité) alors que *AIC* est aucunement consistant mais reste le plus populaire par son ancienneté.

Pour *BIC*, nous avons: $C(N) = \log N$, et pour Φ_β:

$$C(N) = N^{\boldsymbol{\beta}} \log\log N \text{ , avec } 0 < \beta < 1.$$

Notons que le critère *MDL* (*Minimum Description Length*) (Rissanen, 1989), bien connu en théorie de l'information (codage binaire arithmétique des normes de compression), est aussi fortement consistant mais il ne diffère de *BIC* que par des termes négligeables quand N est grand ; c'est pourquoi nous ne conserverons que le critère *BIC*.

Concernant Φ_β, Jouzel et *al.* (1998) ont montré dans que nous avions la condition plus fine :

$$\boldsymbol{\beta}_{\min} = \frac{\log\log N}{\log N} < \beta < 1\text{-}\,\boldsymbol{\beta}_{\min} \text{ ,}$$

pour assurer la consistance forte. Il est aussi montré que $\Phi_{\boldsymbol{\beta}_{\min}}$ est la meilleure valeur de β dans de nombreuses applications.

Enfin, remarquons qu'un avantage de Φ_β est de permettre de généraliser l'écriture des critères : ainsi le cas limite $\beta = 0$ correspond à Φ, alors que β solution de $n^{\beta} \log\log N = \log N$ correspond aux critères *BIC*/*MDL*.

Finalement, l'inégalité : $IC(\theta_i) < IC(\theta_j)$ signifiera que le modèle θ_i réalise un meilleur compromis entre l'adéquation aux données, mesurée par la vraisemblance $L(\theta_i)$, et le coût de ce choix de modèle, mesuré par $|\theta_i|C(N)$. Ainsi θ_i sera préféré au modèle θ_j. Le modèle retenu $\hat{\theta}$ sera donc celui qui minimise le critère *IC* :

$$\hat{\theta} = \arg\min_{\theta_i \in \Theta}[IC(\theta_i)] \tag{2}$$

Ces critères vont être appliqués dans ce papier à la sélection de modèles de lois.

2.2. Génération de séries temporelles synthétiques de rayonnement solaire

Le rayonnement solaire au sol (irradiance globale) peut être représenté comme étant la combinaison de deux composantes : une composante déterministe et une composante stochastique. La composante déterministe représente les variations journalières et saisonnières du rayonnement et peut être décrite par les équations astronomiques bien établies qui décrivent la position du soleil par rapport à la latitude et la longitude du lieu étudié. La composante stochastique est la résultante des phénomènes aléatoires qui affectent le rayonnement solaire comme: la fréquence et la hauteur des nuages, leurs propriétés optiques, la turbidité de l'atmosphère liée à sa composition (contenu en aérosols, vapeur d'eau, ozone, etc.)

La procédure standard pour modéliser une série temporelle de rayonnement synthétique à partir d'une série temporelle de mesures de rayonnement consiste à éliminer les contributions de la composante déterministe afin de rendre la série stationnaire et tenter ensuite de modéliser la série temporelle du terme stochastique. Une fois modélisé le terme stochastique, il suffit de réintroduire la composante déterministe afin d'obtenir une série temporelle synthétique.

Pour isoler la composante stochastique, on emploie soit des méthodes basées sur l'analyse de Fourier (Linares-Rodríguez *et al.*, 2011) (on procède alors par soustraction des contributions fréquentielles de la composante déterministe), soit l'indice de clarté k_t (Graham, 1988 ; Aguiar, 1992 ; Bilbao, 2004 ; Hansen, 2010). L'indice de clarté k_t correspond au rapport entre l'irradiance globale au sol G et l'irradiance globale extraterrestre I_{oh} sur un plan horizontal :

$$k_t = G / I_{oh} \tag{3}$$

$$I_{oh} = I_{sc} E_0 \cos\theta_z \tag{4}$$

avec :

– I_{sc}, l'irradiance produite par la constante solaire ;

– E_0, facteur de correction de l'excentricité ;

– θ_z, angle zénithal solaire.

L'excentricité définit la forme de l'orbite elliptique de la terre autour du soleil ; elle caractérise l'aplatissement de l'ellipse terrestre par rapport à un cercle.

L'angle solaire zénithal, en un lieu donné, est l'angle que fait la droite qui relie le lieu au soleil avec la direction perpendiculaire à la surface du lieu considéré (zénith).

L'indice de clarté, k_t peut être défini pour différents intervalles temporels : horaire, journalier et mensuel. Il peut être aussi calculé avec les composantes directe et diffuse du rayonnement. La conversion du rayonnement en indice de clarté permet de s'affranchir des tendances saisonnières et produit une série stochastique stationnaire. L'indice de clarté permet de comparer les mesures de rayonnement prises à différents instants sans perte d'information sur l'amplitude du rayonnement ; on le note alors $k_t(h)$ où h représente l'heure considérée. Dans le cadre de notre étude nous considérons deux types de modèles.

2.2.1. Le modèle TAG

Le modèle *TAG* (*Time-Dependent Autoregressive Gaussian*), de Aguiar et Collares-Pereira (1992) génère les données synthétiques de rayonnement horaire en utilisant un modèle autorégressif stationnaire, non homogène dans le temps, et suppose une distribution de probabilité gaussienne. Il utilise comme seule entrée la moyenne mensuelle de l'indice de clarté journalier noté K_T. La grande disponibilité de cette donnée moyenne mensuelle un peu partout sur la planète rend ce modèle facilement exploitable. Le modèle *TAG* a pour avantage d'être assez flexible pour modéliser les principales caractéristiques du rayonnement solaire, et assez précis pour être utilisé dans des applications énergétiques. L'étude des propriétés séquentielles du rayonnement solaire par Aguiar et Collares-Pereira a montré qu'il dépend essentiellement de la valeur du rayonnement de l'heure précédente, ce qui les a conduits à proposer le modèle suivant :

$$y(h) = \phi(K_T).y(h\text{-}1) + r_{TAG}(h) \tag{5}$$

C'est un modèle *AR*(1) où $r_{TAG}(h)$ est le bruit blanc dont nous chercherons à identifier la distribution, h est la variable heure, $\phi(K_T)$ est un coefficient de corrélation dépendant de l'indice K_T (voir équation (11)), et $y(h)$ (voir équation (6)) est l'indice de clarté normalisé. La normalisation de l'indice de clarté $k_t(h)$ permet d'obtenir une série temporelle fortement stationnaire. Le modèle ainsi obtenu s'adapte aux données provenant de différents sites de mesures : c'est l'invariance de la loi de probabilité à la localisation.

La normalisation de $k_t(h)$ est réalisée selon l'expression suivante :

$$y(h) = \frac{k_t(h) - k_{tm}(K_T, h)}{\sigma(K_T, h))} \tag{6}$$

où $k_{tm}(K_T, h)$ est la valeur moyenne horaire de $k_t(h)$ et σ est l'écart-type de $k_t(h)$; ils dépendent tous deux de la moyenne mensuelle de l'indice de clarté journalier K_T. $k_{tm}(K_T, h)$ est calculée à partir des paramètres suivants :

$$k_{tm}(h) = \lambda(K_T) + \varepsilon(K_T)(-\frac{\kappa(K_T)}{\sin(h_j)}) \tag{7}$$

$$\lambda(K_T) = -0.19 + 1.12K_T + 0.24\exp(-8K_T) \tag{8}$$

$$\varepsilon(K_T) = 0.32 - 1.6(K_T - 0.5)^2 \tag{9}$$

$$\kappa(K_T) = 0.19 + 2.27{K_T}^2 - 2.51{K_T}^3 \tag{10}$$

$$\phi(K_T) = 0.38 + 0.06\cos(7.4K_T - 2.5) \tag{11}$$

et h_j est l'angle horaire solaire.

2.2.2. Le modèle de Polo

Ce modèle permet de modéliser une série temporelle à partir d'une valeur moyenne et des valeurs aléatoires d'écart type (Polo *et al.*, 2011) Il génère des séries temporelles synthétiques d'indice de clarté k_t à partir des mesures d'indice de clarté moyen, k_{tm}, sur une période donnée. Une des principales conditions imposées par la méthode est que la fréquence et l'amplitude des fluctuations des valeurs de rayonnement générées synthétiquement soient statistiquement représentatives des conditions réelles, c'est-à-dire que la (les) fonction(s) de distribution des données originales soi(en)t comparable(s) à la (aux) fonction(s) de distribution des données générées synthétiquement.

La méthode pour générer les valeurs de rayonnement synthétiques horaires consiste à additionner deux contributions : celle de la moyenne de la période considérée (horaire) et celle de la fluctuation stochastique autour de cette moyenne. Mathématiquement, l'expression du rayonnement synthétique à l'instant *h* peut être formulée ainsi :

$$k(h) = k_{tm}(j) + A(h).\text{sign}(s) \tag{12}$$

où :

– $k_{tm}(j)$ est la valeur moyenne journalière de l'indice de clarté pour le jour *j* ;

– *A(h)* est l'amplitude aléatoire de la fluctuation de l'indice de clarté pour l'heure *h* ;

– *s* est la réalisation d'une distribution gaussienne normale centrée de moyenne nulle et d'écart type unité.

La valeur moyenne horaire du rayonnement peut être obtenue à partir des données mesurées *in situ* et le second terme de l'équation (12) peut être généré selon la procédure suivante :

– Nous générons des valeurs d'écart-type à partir de la distribution identifiée s (par exemple en tirant de façon aléatoire un nombre à partir d'une distribution uniforme et en recherchant la valeur correspondante de la distribution inverse s^{-1}) ;

– Nous multiplions les valeurs d'écart type générées par la valeur maximale des écarts-type mesurés pour obtenir l'amplitude A de la fluctuation ;

– Nous générons un signe aléatoire que nous affectons à l'amplitude A, et le résultat est-ensuite ajouté la valeur moyenne k_{tm}.

Nous cherchons dans cet article à identifier la nature de la distribution du processus A.

3. Identification des lois

3.1. Données utilisées et hypothèses

Pour identifier les distributions de probabilité du bruit blanc r_{TAG} et du processus A, nous utilisons des données horaires de rayonnement fournies par Météo France et mesurées au sol sur deux stations météorologiques localisées à Rochambeau et Ile Royale en Guyane française. Les indices de clarté pour ces deux sites ont été obtenus avec les données horaires de rayonnement extraterrestre calculées au sommet de l'atmosphère. La plage temporelle des données s'étend des années 1996 à 2010, tous les jours de 7 h à 17 h.

Les données mesurées au sol sont en Joules/cm^2. Pour cette phase d'identification des lois, nous choisissons :

– Le site Ile Royale sur 4 années (1998, 2002, 2007,2009)

– Le site Rochambeau sur 7 années (1996, 1998, 2000, 2003, 2005, 2007, 2009).

Nous disposons ainsi de données de l'ordre de $O(10^5)$. De plus, pour le modèle de Polo, l'analyse de la densité de probabilité du processus A est conduite sous différents types de ciel. Nous créons trois classes d'indices de clarté horaires moyens correspondant à trois types de nébulosité du ciel: ciel nuageux, ciel partiellement nuageux, ciel dégagé (Classes C_1 à C_3 respectivement). Les seuils des trois classes ont été choisis afin de disposer d'un nombre de mesures à peu près équivalent dans chacune des classes :

– C_1 : ciel nuageux : $k_{tm} \leq 0.42$;

– C_2 : ciel partiellement nuageux : $0.42<k_{tm}<0.54$;

– C_3 : ciel plutôt dégagé : $k_{tm} \geq 0.54$.

Nous prenons comme hypothèses que les deux processus aléatoires sont indépendants et identiquement distribués (i.i.d.) et que nous pouvons produire une série temporelle :

– de bruits horaires $r_{TAG}(h)$ (par la formule 5),

– et d'écarts types $[k(h) - k_m(j)]$ (par la formule 12). Pour l'obtention plus précise de $A(h)$, nous renvoyons à (Marie-Joseph et *al*, 2013).

Enfin, les lois candidates pour les deux processus sont les suivantes :

– Pour le modèle *TAG*, au vu de la littérature et de l'allure des lois, nous choisissons comme lois candidates les seules lois de Gauss, Logistique et Extreme Value.

– Contrairement au modèle précédent, 9 lois à 1 ou 2 paramètres sont testées pour le modèle de Polo : les lois à 1 paramètre de Rayleigh et exponentielle ; les lois à 2 paramètres Bêta, Gamma, log-normale, gaussienne inverse, de Rice, de Weibull et de Nakagami. Nous renvoyons en Annexe pour l'expression de ces 12 lois.

3.2. Identification

Chaque loi candidate définit un modèle θ_i d'une famille de modèles Θ suivant *TAG* ou Polo. Il s'agit donc de 2 problèmes de sélection de modèles dans lesquels est recherché le meilleur processus A ou r_{TAG} permettant de générer les 2 variables $y(h)$ et $k(h)$. Nous utilisons pour cela les deux critères BIC et $\Phi_{\beta_{\min}}$, le nombre $|\theta_i|$ étant le nombre de paramètres libres de la loi de probabilité considérée des processus A ou rTAG. En réalité, le terme de pénalité n'a une influence que pour le seul modèle de Polo (1 ou 2 paramètres suivant les 9 lois candidates), alors que pour le modèle TAG, les 3 lois candidates ayant 2 paramètres, c'est donc le seul terme de log-vraisemblance $L(\theta_i)$ qui agit.

3.3. Résultats

Nous donnons les résultats d'identification sous forme de tableaux, pour les 2 modèles et leurs lois candidates. Les valeurs indiquées dans les tableaux 1 et 2 sont celles des valeurs moyennes des *IC* : 100 paquets de N = 10 000 données sont considérés et cela pour un bruit observé par ses valeurs horaires (variable h). Les valeurs négatives des critères dans le tableau 1 se justifient par la nature impulsionnelle (à amplitude maximale >> 1) des lois candidates.

Dans le cas du modèle de Polo (tableau 1), à l'issue du test basé *IC*, nous ne retenons que les lois dans le TOP 5 des taux de meilleures reconnaissances. Les critères *BIC* ou $\Phi_{\beta_{\min}}$ ont sélectionné les lois de Nakagami et de Weibull, les lois Bêta, Gamma et exponentielle, quelle que soit la classe de nébulosité. Nous ignorons donc les 4 autres lois dans ce tableau 1.

On remarque que pour les deux modèles, les deux critères *BIC* et $\Phi_{\beta_{\min}}$ donnent les mêmes classements, ce qui est normal compte tenu du nombre élevé de données considérées (N = 10 000), car rappelons que ces deux critères sont consistants (convergence presque sûre). Aucun des deux critères n'offre une meilleure lisibilité que l'autre.

Dans ce tableau 1, suivant la nébulosité, les résultats sont bien différents, ce qui se justifie par la variabilité du modèle suivant l'intensité du rayonnement solaire

(nébulosité). Pour les classes C_1 et C_2, les lois Bêta et de Nakagami sont clairement à distinguer de la loi de Weibull, alors que cette même loi se distingue peu de la loi Gamma (en TOP 2) mais fortement des trois autres candidats dans le cas de faible nébulosité (classe C_3).

Tableau 1. Valeurs moyennes des critères pour les lois du TOP 5, selon les 3 classes de nébulosité, pour le modèle de Polo

classe C_1					
lois / *IC*	***Bêta***	*Weibull*	Exponentielle	Gamma	**Nakagami**
BIC	***-14897***	*-14748*	-14130	-14572	**-14938**
$\Phi_{\beta_{\min}}$	***-14915***	*-14767*	-14139	-14590	**-14957**
classe C_2					
lois / *IC*	**Bêta**	*Weibull*	Exponentielle	Gamma	***Nakagami***
BIC	**-9965**	*-9664*	-8527	-9335	***-9944***
$\Phi_{\beta_{\min}}$	**-9983**	*-9682*	-8537	-9354	***-9962***
classe C_3					
lois / *IC*	Bêta	**Weibull**	Exponentielle	***Gamma***	*Nakagami*
BIC	-11202	**-11577**	-10883	***-11507***	*-11285*
$\Phi_{\beta_{\min}}$	-11220	**-11596**	-10893	***-11526***	*-11595*

Tableau 2. Valeurs moyennes des critères pour les 3 lois candidates, pour le modèle TAG

lois / *IC*	Gaussienne	Logistique	Extreme Value
BIC	***35985***	**28045**	44771
$\Phi_{\beta_{\min}}$	***35973***	**28033**	44759

Dans le tableau 2, les résultats sont encore cohérents concernant les 2 critères d'information. Nous retrouvons bien la loi Gaussienne traditionnellement admise, mais en position 2 de reconnaissance, la loi la plus adéquate, au sens de nos critères de sélection, étant la loi Logistique et la dernière, la loi Extreme Value.

4. Validations des modèles

Pour discuter des conclusions précédentes, nous comparons les densités de probabilité et les fonctions de répartition ou distributions de probabilité cumulative entre le rayonnement solaire mesuré (c'est-à-dire observé) et le rayonnement solaire généré à partir des lois candidates [les variables $x(h)$ ou $k(h)$ suivant le modèle]. À la vue des valeurs des critères d'information obtenues lors de l'étape de sélection de lois, nous ne retenons que les lois retenues en TOP 2. Nous réitérons 500 fois l'expérience sur $N = 10\,000$ échantillons générés et mesurés.

La base des données mesurées pour la validation est évidemment indépendante de celle utilisée pour l'identification (base d'apprentissage) des lois (cf. § 3.1). Ainsi :

– Pour le site de l'Ile Royale : 1999, 2006, 2008, 2010 ;

– Pour le site de Rochambeau : 1997, 1999, 2002, 2004, 2006, 2008 et 2010.

Enfin, les outils de comparaison seront la divergence ou distance de Kullback-Leibler (notée *KL*) entre densités de probabilité, et la distance de Kolmogorov-Smirnov (notée *KS*) entre fonctions de répartition.

Dans les tableaux 3 et 4 nous donnons, pour les modèles de Polo puis *TAG*, les pourcentages de reconnaissances de chacune des 2 lois du TOP 2 suivant les 2 distances considérées (*KL* et *KS*). Le tableau 3 confirme la validité des résultats issus des *IC* dans les cas de nébulosité faible ou forte. Ce n'est pas le cas pour la classe C_2. Nous allons y revenir.

Tableau 3. Pourcentage de reconnaissance par distances KS et KL entre loi mesurée et loi générée. Modèle de Polo

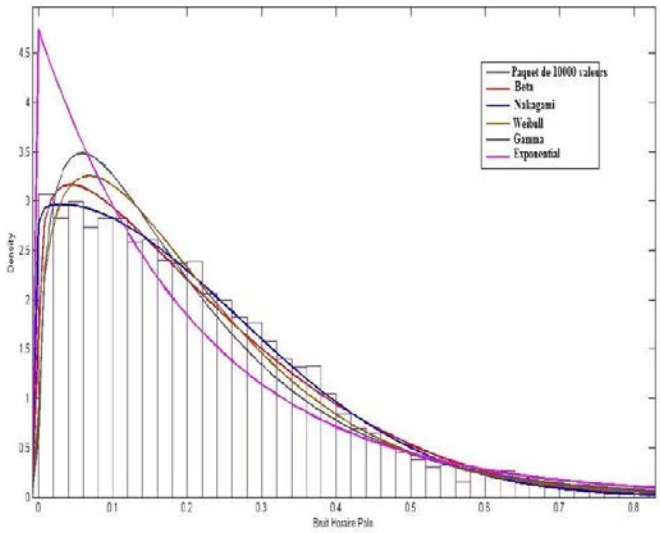
C_i	classe C_1		classe C_2		classe C_3	
lois	Bêta	Nakagami	Bêta	Nakagami	Gamma	**Weibull**
KS	*33%*	**67%**	*37,2%*	**68,2%**	*48%*	**52%**
KL	*33,2%*	**66,8%**	*38,2%*	**61,8%**	*41,2%*	**59,8%**

Tableau 4. Pourcentage de reconnaissance par distances KS et KL entre loi mesurée et loi générée. Modèle TAG

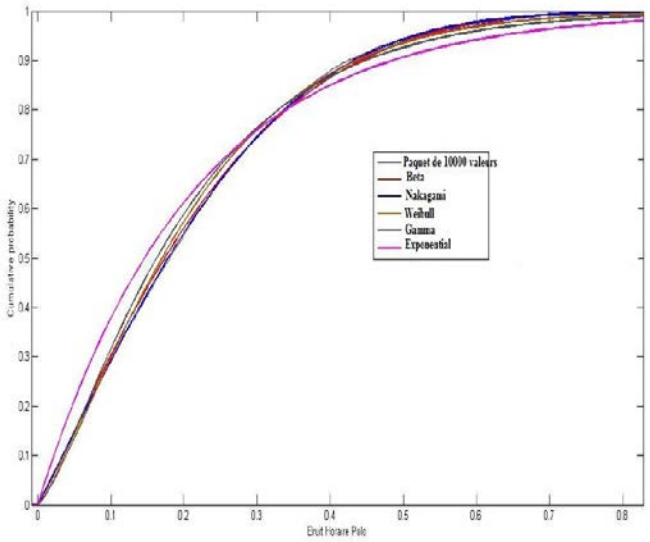
	Gaussienne	Logistique
KS	**35%**	**65%**
KL	**34%**	**66%**

Le tableau 4 confirme les conclusions du tableau 2 dans le cas du modèle *TAG* lorsque nous ne considérons que les lois du TOP 2. On peut donc retenir qu'il vaut

mieux considérer un bruit blanc logistique plutôt que gaussien pour le modèle proposé par Aguiar et Collares-Pereira (1992), ce qui confirmerait les conclusions de Polo (2011).



a) Densité de probabilité



(b) Fonction de répartition

Figure 1. Densités de probabilité de l'irradiance (rayonnement solaire) mesurée, et distributions de probabilité cumulative générées par le modèle de Polo dans le cas C_2

Sur les figures 1 sont indiquées les densités de probabilité (a) et les fonctions cumulatives ou de répartition (b) des 5 lois initialement candidates dans l'hypothèse du modèle de Polo pour le cas litigieux de la classe C_2.

On voit la similitude des courbes et leur comportement proche des valeurs mesurées notamment pour les lois Bêta et de Nakagami et ainsi la difficulté à trancher de manière définitive entre ces deux lois dans ce cas de nébulosité moyenne.

5. Conclusion

Dans cet article nous avons tenté d'améliorer la connaissance des distributions impliquées dans des modèles de génération synthétique du rayonnement solaire en zone tropicale.

L'analyse de onze années de mesures d'irradiance solaire au sol issues de deux stations Météo France situées en Guyane française a été conduite avec une méthode basée sur les critères d'information. Cette méthode d'analyse par critères d'information a permis d'identifier au sens des critères (maximum de vraisemblance pénalisé) les lois de probabilité décrivant le mieux la distribution statistique de chaque terme aléatoire intervenant dans deux modèles de génération de rayonnement solaire synthétique : le modèle *TAG* d'Aguiar et Pereira et le modèle de Polo.

La validation des lois de probabilité identifiées a été menée en comparant des données synthétiques générées durant onze années différentes avec des données réelles et en utilisant la divergence de Kullback-Leibler et la distance de Kolmogorov-Smirnov comme critères de comparaison. Une bonne concordance a été relevée pour le modèle *TAG* d'Aguiar et Pereira entre la loi de probabilité identifiée et la loi de probabilité validée : il s'agit de la loi Logistique. Ce résultat apporte la démonstration du caractère non gaussien du terme aléatoire r_{TAG} du modèle original $AR(1)$, ce qui avait été pressenti par certains auteurs.

Pour le modèle de Polo, une bonne adéquation a été relevée entre les lois de probabilité identifiées et les lois de probabilité validées dans les cas de nébulosité faible et forte ; il s'agit respectivement de la loi de Nakagami et de la loi de Weibull. Dans le cas de nébulosité partielle la proximité des valeurs des critères d'identification ne permet pas de trancher de manière définitive entre ces deux lois tandis que la procédure de validation reconnaît la loi de Nakagami comme étant la loi de distribution des données mesurées.

L'intérêt de la procédure d'identification des lois de distribution présentée dans cet article réside dans le fait qu'elle assure aux modèles de génération synthétique la production de données de rayonnement solaire comparables dans leur contenu statistique aux données mesurées. Autre avantage, cette procédure a permis de mettre en évidence l'invariance dans le temps des lois de distribution représentant les termes aléatoires A et r_{TAG}.

En conclusion, une nouvelle procédure pour déterminer les distributions impliquées dans les termes aléatoires de modèles de rayonnement solaire a été

définie et mise en œuvre avec des résultats concluants. Cette procédure pourrait être étendue à d'autres sites de mesures et appliquée à d'autres modèles de génération synthétique de données solaires horaires ou aussi journalières afin de valider sur une plus grande échelle les conclusions obtenues.

Remerciements

Les auteurs remercient le programme européen FEDER de Guyane et Météo France qui ont permis la réalisation de cette étude dans le cadre du projet de recherche SOLAREST.

Bibliographie

Aguiar R.J., Collares-Pereira M., Conde J.P. (1988). Simple procedure for generating sequences of daily radiation values using a library of Markov transition matrices, *Solar Energy*, vol. 40, n°3, p. 269-279.

Aguiar R.J., Collares-Pereira M. (1992). TAG: A time-dependent, autoregressive, Gaussian model for generating synthetic hourly radiation, *Solar Energy*, vol. 49, n°3, p. 167-174.

Akaike H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, vol. 19, n°6, p. 716-723.

Alata O., Olivier C., Pousset Y. (2013). Law recognitions by information criteria for the statistical modeling of small scale fading of the radio mobile channel. *Signal Processing*, vol. 93, n°5, p. 1064-1078.

Amado M., Poggi F. (2012). Towards Solar Urban Planning: A New Step for Better Energy Performance, *Energy Procedia*, vol. 30, p. 1261-1273.

Bilbao J., Miguel A., Franco J.A. and Ayuso A. (2004). Test Reference Year Generation and Evaluation Methods in the Continental Mediterranean Area. *Journal of Applied Meteorology*, vol. 43, p. 390-400.

Boland J. (1995). Time Series Analysis of Climatic variables, *Solar Energy*, vol. 55, n°5, p. 377-388.

Boland J. (2008). Time Series Modelling of Solar Radiation. *Modeling Solar Radiation at the Earth Surface – Recent Advances.* Viorel Badescu Ed. Springer Verlag, Chapter 11.

El Matouat A. and Hallin M. (1996). Order selection, stochastic complexity and Kullback-Leibler information. *Time Series Analysis*, Springer Verlag, 2: 291-299.

Graham V.A., Hollands K.G.T., Unny T.E. (1988). A time series model for K_t with application to global synthetic weather generation, *Solar Energy*, vol. 40, n°2, p. 83-92.

Hannan E.J., Quinn B.G. (1979). The Determination of the Order of an Autoregression, *Journal of the Royal Statistic Society*, vol. 41, n°2, p. 190-195.

Hansen C.W., Stein J.S. and Ellis A. (2010). *Statistical Criteria for Characterizing Irradiance Time Series.* Sandia Report, SAND2010-7314, October 2010.

Jouzel F., Olivier C., El Matouat A. (1998). Information Criteria based edge Detection, *EUSIPCO'98-Signal Processing IX,* Rhodes (Greece), vol. 2, p. 997-1000, Sept 1998.

Linares-Rodríguez A., Antonio Ruiz-Arias J., Pozo-Vázquez D., Tovar-Pescador J. (2011). Generation of synthetic daily global solar radiation data based on ERA-Interim reanalysis and artificial neural networks, *Energy*, vol. 36, n°8, p. 5356-5365.

Marie-Joseph I., Linguet L., Gobinddass M.-L., Wald L. (2013). On the applicability of the Heliosat-2 method to assess surface solar irradiance in the Intertropical Convergence Zone, French Guiana, *International Journal of Remote Sensing*, vol. 34, n° 8, p. 3012-3027.

Mellit A., Kalogirou S.A., Shaari S., Salhi H., Hadj A. (2008). Methodology for predicting sequences of mean monthly clearness index and daily solar radiation data in remote areas: Application for sizing a stand-alone PV system, *Renewable Energy*, vol. 33, n°7, p. 1570-1590.

Muselli M., Poggi P., Notton G. Louche A. (1998). Improved procedure for stand-alone photovoltaic systems sizing using meteostat satellite images. *Solar Energy*, vol. 62, p. 429-444.

Olivier C., Alata O. (2009). The Information Criteria: examples of applications in image and signal processing. *Optimisation in Image and Signal Processing*, Wiley Ed., Chapter 4, p. 79-110.

Poggi P., Notton G., Muselli M. Louche A. (2000). Stochastic study of hourly total solar radiation in Corsica using a Markov model. *International Journal of Climatology*, vol. 20, p. 1843-1860.

Polo J., Zarzalejo L.F., Marchante R. and Navarro A.A. (2011). A simple approach to the synthetic generation of solar irradiance time series with high temporal resolution, *Solar Energy*, vol. 85, p. 1164-1170.

Remund J. and Page J. (2002). *Integration and exploitation of networked Solar radiation Databases for environment monitoring*, SODA Project Report.

Rissanen J. (1989). Stochastic Complexity in Statistical Inquiry, *World Scientific ed*., New Jersey.

Schwarz G. (1978). Estimating the dimension of a model. *The Annals of Statistics,* vol. 6, p. 461-464.

Stanhill G., Cohen S. (2001). Global dimming: a review of the evidence for a widespread and significant reduction in global radiation with discussion of its probable causes and possible agricultural consequences. *Agricultural and Forest Meteorology*, vol. 107, n° 4, p. 255-278.

Tiba C., Fraidenraich N. (2004). Analysis of monthly time series of solar radiation and sunshine hours in tropical climates. *Renewable Energy*, vol. 29, n°7, p. 1147-1160.

Annexe

Nous rappelons ici les densités de probabilité des différentes lois candidates utilisées dans ce papier.

1. Pour le modèle *TAG,* 3 lois sont considérées :

– La loi Gaussienne

$$f(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in R$$

avec μ, la moyenne de la distribution, et σ l'écart-type de la distribution.

– La loi Extreme Value

$$f(x,\mu,\sigma) = \sigma^{-1} e^{\frac{(x-\mu)}{\sigma}} e^{-e^{\frac{(x-\mu)}{\sigma}}}, \quad x \in R$$

avec μ et σ, les paramètres de forme de la distribution.

– La loi Logistique

$$f(x,\mu,s) = \frac{1}{s\left(1+e^{\frac{x-\mu}{s}}\right)^2} e^{-\frac{(x-\mu)}{s}}, \quad x \in R$$

avec μ la moyenne et s un paramètre de forme lié à la variance.

2. Pour le modèle de Polo, 9 lois sont considérées:

– La loi Bêta

$$f(x,\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 y^{\alpha-1}(1-y)^{\beta-1}dy} \quad x \in [0,1]$$

= 0 sinon

avec α et β, les paramètres de forme de la distribution.

– La loi Gamma

$$f(x,\alpha,\beta) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1}\beta^{\alpha}e^{-\beta x} \quad x \geq 0$$

= 0 sinon

avec :

- α un paramètre de forme de la distribution et β un paramètre d'intensité,
- $\Gamma(\alpha)$, la fonction Gamma.

– La loi Exponentielle

$$f(x,\mu) = \frac{1}{\mu} e^{-\frac{x}{\mu}}, \quad x \geq 0$$

= 0 sinon

avec μ la moyenne de la distribution.

– La loi de Nakagami

$$f(x,m,\omega) = 2\left(\frac{m}{\omega}\right)^m \frac{1}{\Gamma(m)} x^{2m-1} e^{-\frac{m}{\omega}x^2}, \quad x > 0$$

= 0 sinon,

avec :

- m un paramètre de forme, ω un paramètre permettant de contrôler la propagation de la distribution,
- $\Gamma(m)$ la fonction Gamma.

– La loi de Rayleigh

$$f(x,m) = \frac{x}{m^2} e^{-\frac{x^2}{2m^2}} \qquad x \geq 0$$

= 0 sinon,

avec : m le mode de la distribution.

– La loi de Rice

$$f(x,s,\sigma) = I_0\left(\frac{xs}{\sigma^2}\right)^m \frac{x}{\sigma^2} e^{-\left(\frac{x^2+s^2}{2\sigma^2}\right)}, \quad x \geq 0$$

= 0 sinon,

avec :

- s et σ, les paramètres de forme de la distribution
- I_0, la fonction Bessel de 1ère espèce d'ordre 0.

– La loi de Weibull à 2 paramètres

$$f(x,s,\sigma) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}, \quad x \geq 0$$

= 0 sinon,

avec k un paramètre de forme et λ un paramètre d'échelle.

– La loi Gaussienne inverse

$$f(x,\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} e^{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}}, \quad x > 0$$

$$= 0 \text{ sinon,}$$

avec μ la moyenne et λ un paramètre de forme.

– La loi log-normale

$$f(x,\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, \quad x > 0$$

$$= 0 \text{ sinon,}$$

avec μ la moyenne et σ l'écart type du logarithme de la variable.