Classification of daily solar radiation distributions using a mixture of Dirichlet distributions.

Ted Soubdhan, Richard Emilion, Rudy Calif

Classification of daily solar radiation distributions using a mixture of Dirichlet distributions.

Ted SOUBDHAN*[1], **Richard EMILION**[2] **and Rudy CALIF**[1]

[1]GRER
Université des Antilles et de la Guyane
BP 592 UFR SEN - 97159 Pointe à Pitre, Guadeloupe

[2] MAPMO, UMR CNRS 6628 Université d'Orléans
B.P. 6759 45067 Orléans cedex 2, France

Abstract :

In order to qualify the fluctuating nature of solar radiation under tropical climate, we classify daily distributions of the clearness index k_t by estimating a finite mixture of Dirichlet distributions without assuming any parametric hypothesis on these daily distributions. The method is applied to solar radiation measurements performed in Guadeloupe (16°2 N, 61 W) where important fluctuations can be observed even within a short period of a few minutes. The results put in evidence four distinct classes of distributions corresponding to different types of days. The sequence of such classes can be of interest for prediction.

Keywords: clearness index k_t; daily distributions; Dirichlet mixture; classification; solar radiation; solar systems

1. Introduction

Under tropical climate solar radiation is a very fluctuating data, notably because of numerous clouds. Fast changes in the local meteorological conditions, as observed in tropical climate can provoke large variation of solar radiation. The amplitude of these variations can reach 700 W/m² and occur within a short time interval, from few seconds to few minutes according to the geographical location. These variations depend on the clouds size, speed and number.

*corresponding author: ted.soubdhan@univ-ag.fr

Studies of solar energy systems are traditionally performed using daily or hourly data [1], [2]. Such data do not take into account the fluctuations mentioned previously although it has been shown that the fractional time distribution for instantaneous radiation differs significantly from that obtained with daily values [3] [4].

Rapid variations of solar energy induce rapid and large variations of the output of solar systems, such as photovoltaic solar cells (PV) used for electrical production, due to their very short response time [5]. Therefore in power distribution grids with high density of PV, rapid fluctuations of the produced electrical power can appear, leading to unpredictable variations of node voltage and power in electric networks. In small grids as they exist on islands (such as in Guadeloupe, FWI) such fluctuations can cause instabilities.

Hence management of the electrical network and of the alternative power sources requires a better identification of these small time scales variations. This stresses the need for power system operators to develop real time estimation tools for such disturbances in order to anticipate power shortages and power surges.

In this paper, we first summarize the variations of each daily solar radiation by an histogram which estimates the distribution of the daily solar radiation from measurements performed in Guadeloupe with a sampling rate of 1 Hz. Then we classify these histograms by estimating a finite mixture of Dirichlet distribution. This yields four classes corresponding to four types of days. The sequence of the classes can be of interest for prediction.

The paper is organized as follows. Section 2concerns the experimental set-up of solar radiation measurements. We present our method for creating empirical histograms from the measurements in section 3. In section 4, we present an overview of solar radiation classification and present the main ideas of our method. In section 5 and 6 we present the theoretical framework of the method. In section 7 and 8 we present the classification results and an analysis of the sequence of classes. In section 9 we discuss of the interest of such a classification for PV systems and we conclude in section 10.

2. Solar global radiation measurements

Our global solar radiations measurements were performed in Guadeloupe, an island in the West Indies which is located at 16°15 N latitude and 60° 30 W longitudes. In such a tropical zone, solar radiation is an important climatic data to be taken into account as the daily average for the solar load for a horizontal surface is around 5 kWh/m². A constant sunshine combined with the thermal inertia of the ocean makes the air temperature variation quite weak, between 17°C and 33°C with an average of 25°C to 26°C. Relative humidity ranges from 70% to 80% and the trade winds are relatively constant all along the year.

Our measurements sampled at 1 Hz were performed during one year, from September 2005 to December 2006. Only few authors have performed measurements with such a time step [3], [6], [7]. The measurements made with two pyranometers from KIPP&ZONEN were recorded by a CAMPBELL SCIENTIFIC data logger.

On Figure 1 below, are presented two examples of measurements and their corresponding pdf.

3. Data representation

We consider a sample $(h_1,..., h_i,...,h_n)$ of $n=365$ daily histograms of k_t based on a fixed partition of the clearness index range into $l=20$ intervals of equal length. Each histogram h_i of k_t has therefore l bins, say $h_i = (h_{i,1},...,h_{i,l})$, such that $h_{i,j} \geq 0$ and $\sum_{j=1}^{l} h_{i,j} = 1$.

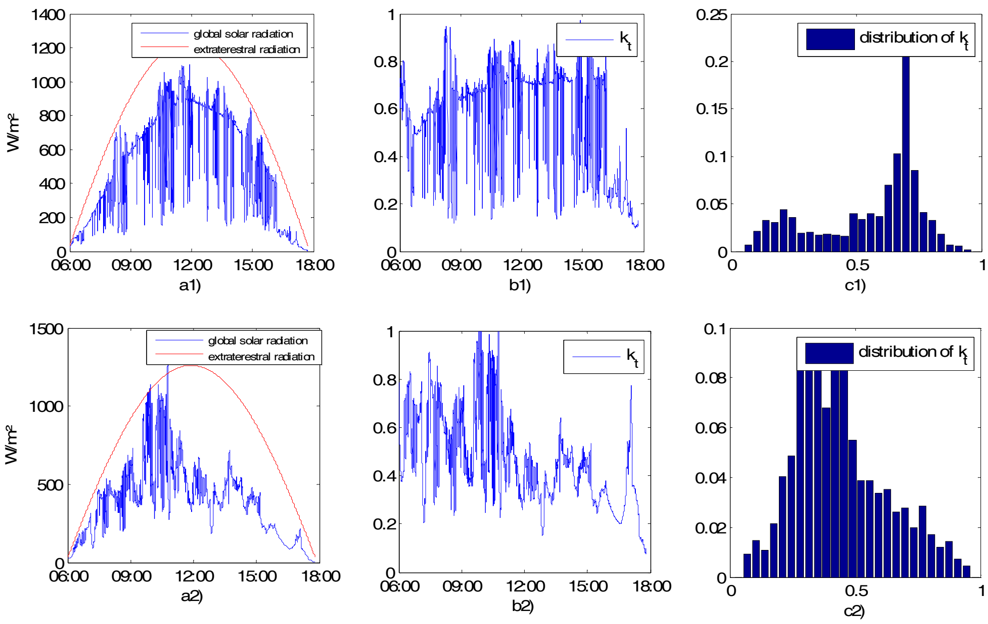

Figure 1: Two examples of day: daily radiation (a), clearness index kt, (b) and corresponding histogram (c).

In the sequel it will be important to consider that *conditionally to the observed data*, $h_i = (h_{i,1},...,h_{i,l})$ is a probability vector on the finite set $V=\{1,\ldots,l\}$.

The choice *l=20* was decided empirically as a compromise between accurate estimation of the distribution and getting nonempty bins.

4. Daily solar radiation classification

4.1 An overview of solar radiation classification

Classification of days based on daily solar radiation properties has been investigated in many studies, generally using a supervised and a parametric approach.

For example, Boullier and Le Chapellier [9] present a classification based on twelve parameters (horizontal irradiation, air temperature, wind speed, humidity, nebulosity, and so on) which allow a study of the transitions between diurnal and nocturnal periods. They get 4 groups (9 sub-classes) for “day” periods and also 4 groups (9 sub-classes) for “night”' periods by combining the discriminant character of each variable.

Fabero et al. [10] put in evidence nine “typical” day periods by decomposing each studied day in three periods. This study, performed in Madrid (Spain), use global horizontal irradiation recorded at 1/600 Hz to point out transitions within a day in order to evaluate a solar potential.

Muselli et al [11] propose a classification methodology with parameters defined from hourly clearness index profiles. A Ward aggregation classification method is then applied on these parameters. This study done with a monthly (resp. an annual) time step leads to 3, 4 or 5 classes (resp. 3 classes).

Maafi et al [12] and Harrouni et al [13] use fractal dimension and daily clearness index k_t as classification parameters. Three classes obtained by some specific thresholds of these parameters correspond to clear sky, partially clouded and overcast sky, respectively.

4.2 Main ideas of our classification method

Our aim is to apply statistical inference techniques for identifying classes of daily k_t histograms as well as for evaluating the probability that a daily k_t histogram belongs to each of these classes. We proceed as follows.

First we consider a sample of daily k_t histograms obtained from one second measurements. Such histograms are nonparametric estimators of distributions so that they incorporate many aspects of a day behaviour, such as all order moments or tail behaviour.

Next a classification algorithm is performed on the sample of histograms so that we get classes of distributions. The variability of the distributions within a specific class is represented by a random distribution which depends on this class. A very popular example of random distribution is the Dirichlet distribution which is suitable for representing a wide variety of distributions. The variability of the whole sample will be represented by a mixture of Dirichlet distributions.

To estimate the parameters of the mixture as well as the parameters of each Dirichlet distribution we use the SAEM algorithm which is a stochastic version improving the celebrated EM algorithm.

The above method is interesting not only for capturing the entire range of daily solar radiation behaviour with all its statistical characteristics, and also for dealing with almost any type of distribution but also for describing a class by a random distribution. This last point is much more general and interesting than describing classes by tuples (e.g., mean-variance) or by thresholds, as seen in the overview section.

Note that this method has been applied for Internet traffic flow classification [14]. A contribution of the present paper is to study the applicability of this method for daily solar radiation classification.

5. Statistical setting

5.1 Random Distribution

A random distribution (RD) is a measurable map from a probability space $(\Omega, \mathcal{F}, P)$ to the space $P(V)$ of all probability measures defined on a fixed measurable set $(V, \mathcal{V})$.

If $X : \Omega \to P(V)$ is a RD, its distribution P_X is then a probability measure on $P(V)$.

If $V = \{1, \ldots, l\}$ is the above finite set, then note that $P(V)$, the set of all probability measures defined over V, can be identified to be the set

$\{x = (x_1, \ldots, x_l)$ such that $x_j \geq 0$ and $\sum_{j=1}^{l} x_j = 1\}$

A popular example of RD is the following one.

5.2 Dirichlet distribution and Dirichlet density

Consider l independent random variables $Z_1, \ldots, Z_l$ following a gamma distribution $\gamma(\alpha_1, 1)$, ..., $\gamma(\alpha_l, 1)$ respectively, where $\gamma(a, b)(x) = \frac{1}{\Gamma(a)}(1-x)b^a e^{b-x} x^{a-1}\ I_{(0,+\infty)}(x)dx$.

If we normalize each random variable Z_l by the sum $Z = Z_1 + \dots + Z_l$, then the distribution of the random vector $X = (\frac{Z_1}{Z},\dots,\frac{Z_l}{Z})$ is called the Dirichlet distribution $\mathcal{D}(\alpha_1,\dots,\alpha_L)$.

Observe that since $\frac{Z_j}{Z} \geq 0$ and $\sum_{j=1}^{l}\frac{Z_j}{Z} = 1$, X is a random distribution on $P(V)$, $V = \{1 \dots, l\}$.

Hence the Dirichlet distribution is a first natural example of distribution of a random distribution, that is a probability measure on $P(V)$, $V = \{1,\dots,l\}$.

It can be shown that the *l-1*dimensionnal random vector $(\frac{Z_1}{Z},\dots,\frac{Z_{l-1}}{Z})$ has the following density with respect to the Lebesgue measure on the set:

$T_l=\{x = (x_1, \dots, x_{l-1})$ such that $x_j \geq 0$ and $\sum_{j=1}^{l-1} x_j \leq 1\}$,

the so-called Dirichlet density:

$$f(x_1, x_2, \dots, x_{l-1} \,\Big|\, \alpha_1, \alpha_2, \dots, \alpha_l) = \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_l)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\dots\Gamma(\alpha_l)} \prod_{i=1}^{l-1} x_j^{\alpha_j - 1} (1 - \sum_{i=1}^{l-1} x_j)^{\alpha_l - 1}$$

for $(x_1, \dots, x_{l-1}) \in T_l$

5.3 Mixture of Dirichlet distributions

Let $X : \Omega \to P(V)$ be a RD, with, as above, $V = \{1, \dots, l\}$ and

$P(V)=\{ x = (x_1, \dots, x_l)$ such that $x_j \geq 0$ and $\sum_{j=1}^{l} x_j = 1\}$.

The distribution of P_X of X is a finite mixture of Dirichlet distributions if it is a convex combination of K standard Dirichlet distributions $D(\alpha_1^k,\dots,\alpha_l^k)$:

$$P_X = \sum_{k=1}^{K} p_k D(\alpha_1^k,\dots,\alpha_l^k) \text{ with } p_k > 0 \text{ and } \sum_{k=1}^{K} p_k = 1$$

The mixture problem consists in estimating the parameters p_k and $(\alpha_1^k,\dots,\alpha_l^k)_{1\leq k\leq K}$.

To achieve this aim we use an iterative algorithm where the inputs are the n histograms vectors $h_i = (h_{i,1},\dots,h_{i,l})$, $i=1,2,\dots,n$ as described before and the number of classes, K:

5.4 Dirichlet distribution properties

The Dirichlet distribution has the following nice property that is particularly useful. If $X=(X_1, ..., X_l)$ has a Dirichlet distribution, $\mathcal{D}(\alpha_1,...,\alpha_l)$, then the marginal distribution of each component X_l follows a beta distribution :

$$X_l \sim \mathcal{B}(\alpha_l, A-\alpha_l)$$

where A = $\sum_{l=1}^{L} \alpha_l$ is the massvalue.

Recall that the Beta distribution $\mathcal{B}(\alpha,\beta)$ is defined by the probability density function on [0;1]:

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(1-x)^{\beta-1}x^{\alpha-1}$$

These properties make the Dirichlet distribution very attractive for modelling random distributions. Moreover Dirichlet distributions, and more specially the mixtures of Dirichlet distributions are very suitable to encompass a very large spectrum of various distributions appearing in the real world [15, 6, 7].

6. Estimation procedure

In the present work, as said before, we want to classify daily pdfs of k_t based on the similarity of their distribution. We assume that the observed empirical histograms are coming from a source governed by a random distribution. Rather than finding a single distribution to represent all the histograms, it makes intuitive sense to think that each class of days is represented by a specific Dirichlet distribution while the entire ensemble of days is represented by a finite mixture of K Dirichlet distributions $\sum_{k=1}^{K} p_k D(\alpha_1^k,...,\alpha_l^k)$ where component $D(\alpha_1^k,...,\alpha_l^k)$ describes class k, and p_k is its weight.

However in practice these parameters are unknown and in order to fit our model, we need to estimate them.

6.1 Algorithm

The point is to consider the above daily k_t histograms as iid, (independent and identically distributed) outcomes of a random distribution X having for distribution a finite mixture of Dirichlet distributions. The number K of components of the mixture will be fixed by the user. The following SAEM procedure is a stochastic variant of the EM alogorithm.

Description of the algorithm:

Initialization step: Assign randomly each histogram h_i, $i = 1,...,n$ to a class

Simulation step: Generate randomly $t_{i,k}^{(0)} (i = 1,...,n)$ representing the initial a posteriori probability that a histogram *i* is in a class *K* where $1 \le k \le K$

For *q=0* to *Q* **do**

Stochastic step: Generate random multinomial numbers $e_{qi} = (e_{qi}^k)$ following the probability distribution $\{ t_{i,k}^{(q)} \}$ where all the $e_{qi}^{(k)}$ are 0 except one of them equal to 1. We then get a partition $C = (C_k)_{k=1,...,K}$ of the set of histograms

If $\frac{\sum_{i=1,...,N} e_{qi}^k}{N} < c(n)$ for some *k*

then return back to the initialisation step.

Maximisation step: Estimate the mixing weights:

$$p_k^{(q)} = \frac{1}{n}[(1-\gamma_q) \sum_{i=1,...,N} t_{ik}^q + \gamma_q \sum_{i=1,...,N} e_{qi}^k]$$

and the parameter value:

$$\alpha_k^{(q)} = (1-\gamma_q) \frac{\sum_{i=1,...,N} t_{ik}^q b(f_i)}{\sum_{i=1,...,N} t_{ik}^q} + \gamma_q \frac{\sum_{i=1,...,N} e_{qi}^k b(f_i)}{\sum_{i=1,...,N} e_{qi}^k}$$

Estimation step: Update the a posteriori probability of a histogram *i* belongs to class k according to:

$$t_{ik}^{(q+1)} = \frac{p_{q+1,k} D(\alpha_{1,q+1}^k,...,\alpha_{l,q+1}^k)(y_i)}{\sum_{k=1,...,K} p_{k,q+1} D(\alpha_{1,q+1}^k,...,\alpha_{l,q+1}^k)(y_i)}$$

End For

7. Results

The proposed algorithm is applied on the whole set of *n=365* daily histograms of k_t in order to find *K* classes which is also the number of components in the Dirichlet mixture model.

The above algorithm has been tested with *K=2,3,4,5,…*. With the dataset used here, the most suitable value is *K=4,* which corresponds to the following 4 different representative types of solar radiation days: Clear sky days, Intermittent clear sky days, Cloudy sky days, Intermittent cloudy sky days.

Let us define 3 characteristics of a day:

Sunshine (*S*) with *S=1* for an important solar radiation and *S=0* otherwise.

Cloudy level (*C*) with *C=0, 1, 2.*

Dynamic level (*D*) with *D=0, 1 , 2* corresponding to the solar radiation dynamic due to cloud sizes and cloud speed (high *D* corresponding to frequent passages).

Class 1: Clear sky days (S=1, C=0, D=0)

The first class is composed of monomodal distributions of k_t having a maximum occurrence value around $k_t = 0.75$ (Figure *2*). These distributions are representative of clear sky conditions, that is days of solar radiation with very few clouds and thus a very slow dynamic, as shown in Figure *3*. This can be observed on the light tailed distributions where the pdf of k_t is less than *0.1* on the range $k_t \in [0; 0.6]$. The weight of this small class of *30* days is *8%*.

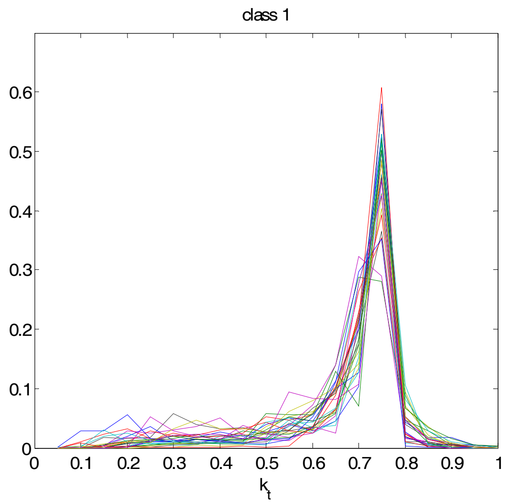

Figure 2: daily distributions of k_t in class *1*

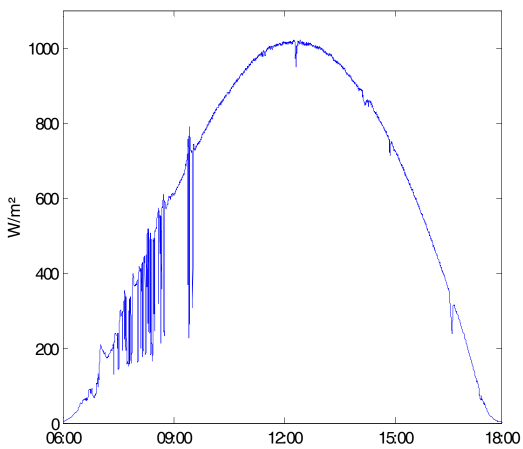

Figure 3: A clear sky day

Class 2: Intermittent clear sky days (S=1, C=1, D=1)

The second class is composed of monomodal distributions of k_t having a maximum around $k_t = 0.75$ (Figure *4*) but with heavier tails compared to distributions of class *1*. This is representative of days with an important solar radiation with some clouds corresponding to a medium level dynamic as shown in Figure *5*. The weight of this class of *128* days is *35%*.

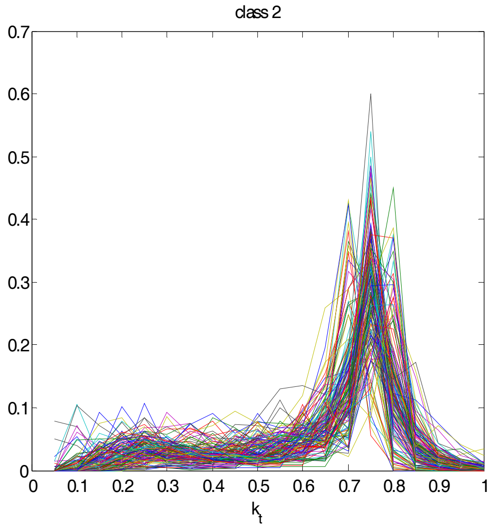

Figure 4: Daily distributions of kt in class 2

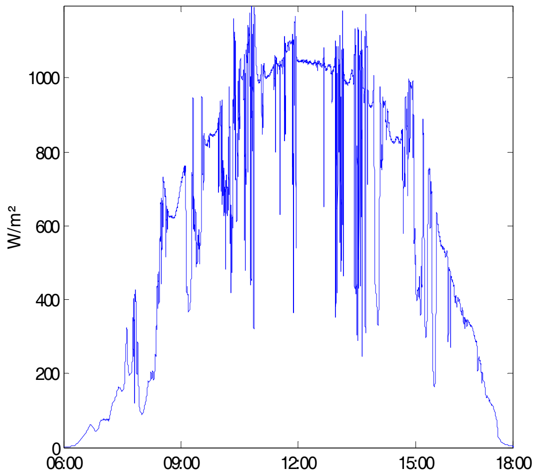

Figure 5: A day of class 2

Class 3: Cloudy sky days (S=0, C=2, D=0)

The third class is composed of monomodal distributions of k_t having a maximum value for *k_t=0.1* (Figure *6*). These distributions are representative of completely cloudy sky days with big size clouds having a slow speed so that the dynamical level is weak (see figure 7). In that case the solar radiation is mainly diffused by clouds. The weight of this class of *25* days is *7%*.

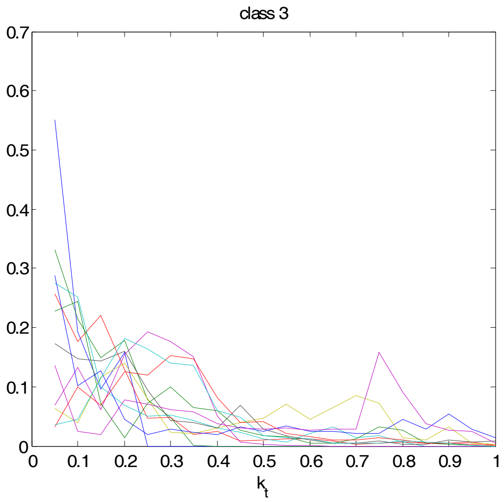

Figure 6: Daily distributions of kt in class 3

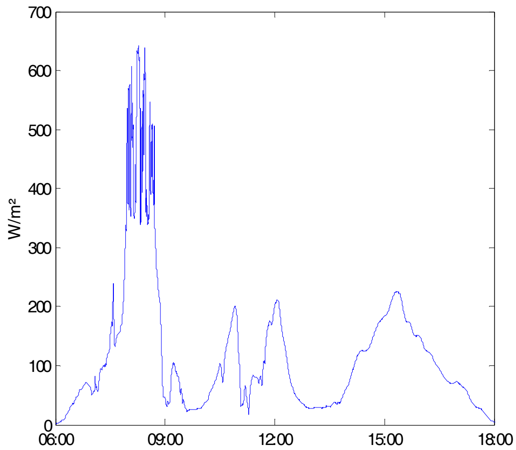

Figure 7: A day of class 3

Class 4: Intermittent cloudy days (S=1, C=2, D=2)

The fourth class is composed of bimodal distributions of k_t. They have two maximum value, one around k_t *=0.25* and the other one around k_t *=0.75* (Figure *8*). These distributions are representative of days with an important sunshine combined with a large number of small clouds with high speed of passages and thus with high dynamic level (Figure *9*). In this class we have *190* days that is *52%* of the measured data .

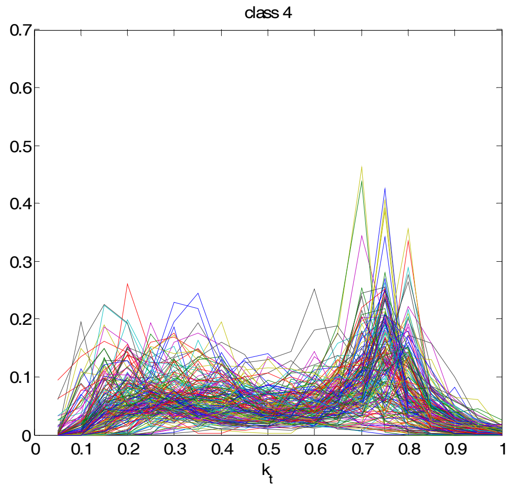

Figure 8: Daily distributions of kt in class 4

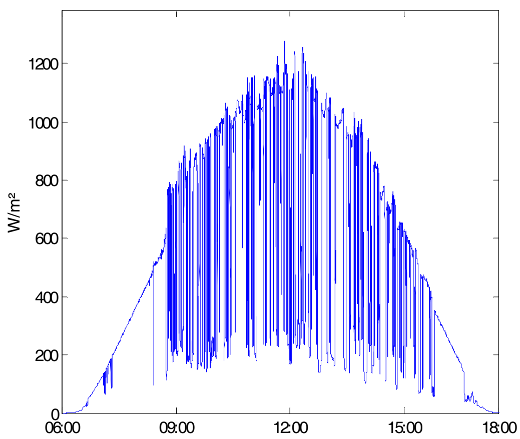

Figure 9: A day of solar radiation in class 4

Mean pdf and weight of each class

Each class mean PDF and weight is plotted in figure *10*, respectively in figure *11*.

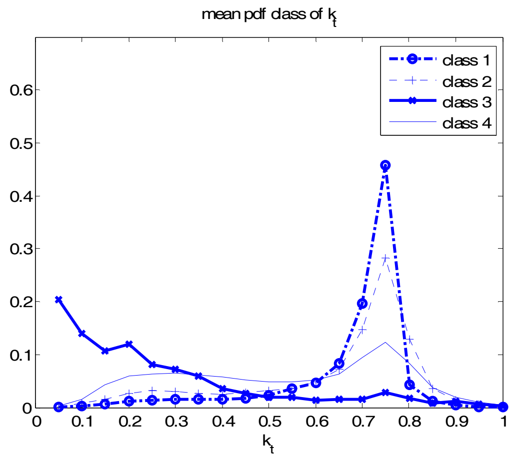

Figure 10: Mean pdf of kt for each class

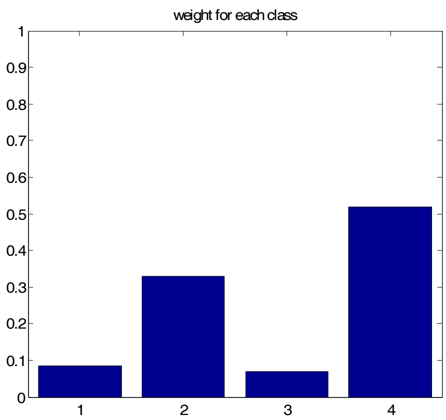

Figure 11: Weight of each class

According to figure 11, days of class 1 and class 3 are in minority as they correspond to rare meteorological events. Days of class 3 which are intermittent cloudy, form the major component of our dataset.

8. Sequence of Classes

Once the *n=365* days are classified into *4* classes, each day can be replaced by its class number, so that we obtain a *{1, 2, 3, 4}*-valued sequence of length *n=365* as plotted in figure 12. This can represent the yearly evolution of the type of solar radiation days.

It can be observed that this sequence has some interesting statistical properties such as an exponential residence time distribution in each class (figure 13) and also the transition from a class to another.

This leads us to think that such a sequence can be a path of a discrete Markov chain or a Hidden Markov Chain Model having 4 states *{1, 2, 3, 4}*. This can be of interest for further research on solar radiation prediction.

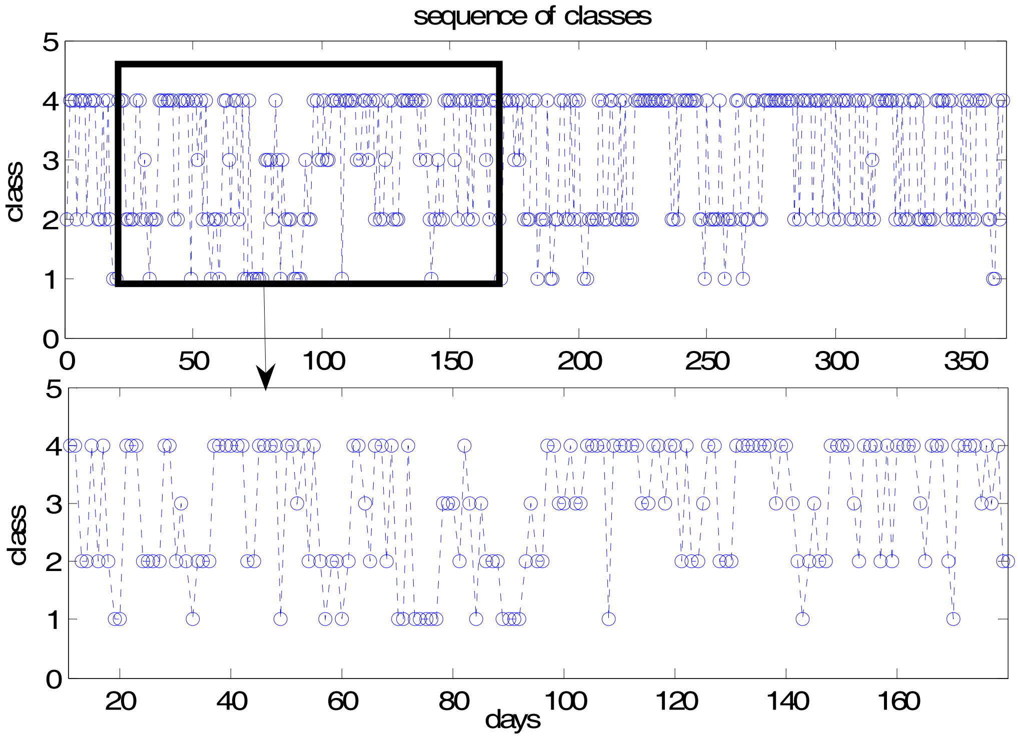

Figure 12: sequence of classes

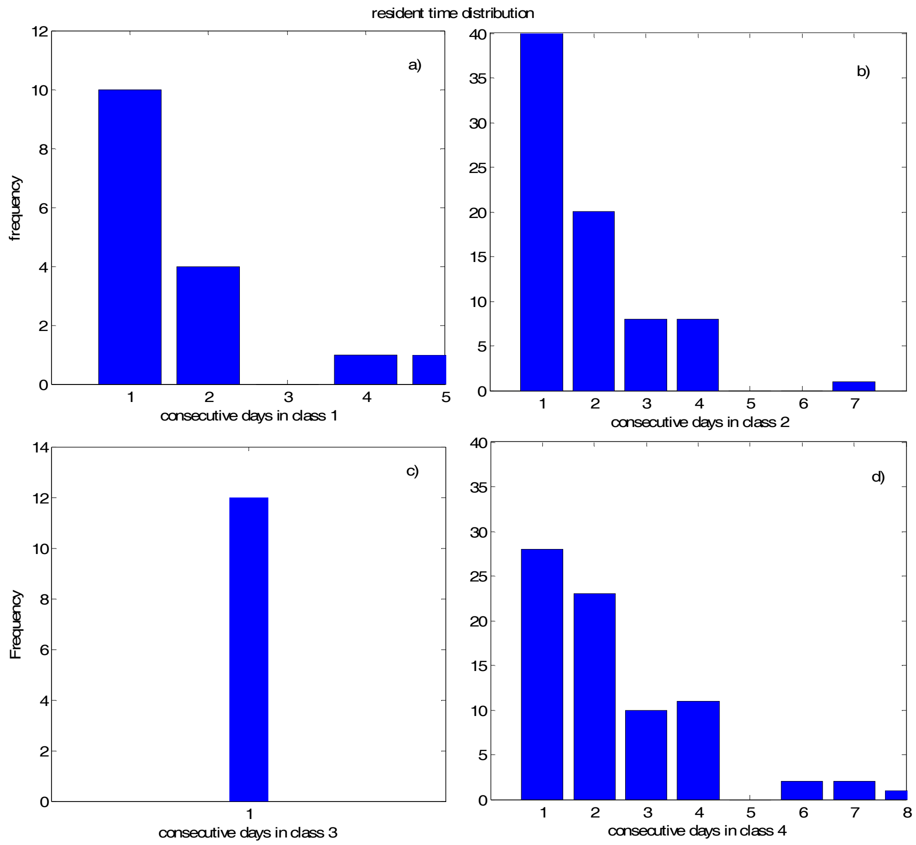

Figure 13: resident time distribution for each class

9. Interest for solar PV systems

As said in the introduction, a good knowledge of solar radiation and its dynamic is of major interest for solar energy systems particularly for PV systems.

In the studied dataset we observe a majority (52%) of intermittent cloudy days (Class 4) with a high solar radiation level and high dynamic due to small clouds passages. In spite of high solar radiation level, such days can cause instabilities in an electrical network with PV systems because of high fluctuations on the PV output.

Class 1, with low dynamic and high level of sunshine, is very suitable for electricity production with PV, but for the given site it only represents 10 days over a year.

Class 2 (35% of days), with a high level of sunshine and medium dynamic level is also suitable for PV systems.

Some interesting information can be observed from the sequence of classes. While class 2 and class 4 appear all along the year, class 1 do not appear from October to December and class 3 mainly occur from March to June. This suggests the existence of various regimes in solar radiation and leads us to think that Hidden Markov Model (HMM) should be a more suitable model.

Finally observe that duration of class *3* is always 1 day (figure 13-c)).

10. Conclusion

We have first summarized the daily solar radiation fluctuations under tropical climate by the distributions of the clearness index k_t, estimated by histograms having $l = 20$ bins. Then, conditionally to the measurements data, these histograms have been classified by estimating a finite mixture of Dirichlet distributions without assuming any parametric hypothesis on the daily distributions.

It has been proved in [15] [16] that if the sampling rate of the measurement is higher and the histograms are refined with a larger number of bins then the classification procedure is consistent, that is the classes will remain quite identical.

The method has been applied to solar radiation measurements performed in Guadeloupe (16°2 N, 61 W) where important fluctuations can be observed even within a short period of a few minutes. The results put in evidence four different classes of distributions corresponding to different types of days:

- Clear sky days, with high level of sunshine, very few clouds and thus low dynamic;
- Intermittent clear sky days, with high level of sunshine, small clouds and medium dynamic;
- Cloudy sky days, with low level of sunshine, big size clouds and low dynamic
- Intermittent cloudy sky days, with high level of sunshine, high number of small clouds, and high dynamic.

Classes 1 and 3 are small (*8%* and *7%* respectively) while classes 2 and 4 are large (*35%* and *52%* respectively).

Analyzing the time sequence of classes leads us to think that the solar radiation days are governed by a Hidden Markov Chain having 4 states $\{1, 2, 3, 4\}$ with some underlying unobservable *regimes* of solar radiation. This can be of interest for further researches, notably for the prediction of a specific class of days.

11. References

[1]. Notton, G., Cristofari. C., Poggi P., Muselli M., Calculation of solar irradiance profiles from hourly data to simulate energy systems behaviour. Renewable Energy 27 123–142, 2002.

[2]. Suehrcke, H., McCormick, P.G., 1988. The frequency distribution of instantaneous insolation values. Solar Energy 40, 413–422.

[3]. Tovar, J., Olmo, F.J., Alados-Arboledas, L., 1998. One-minute global irradiance probability density distributions conditioned to the optical air mass. Solar Energy 62, 387–393.

[4]. Woyte A., Belmans R., Nijs J. Fluctuations in instantaneous clearness index: Analysis and statistics (2007) Solar Energy, 81 (2), pp. 195-206

[5]. Woyte, A., Thong, V.V., Belmans, R., Nijs, J.c., Voltage fluctuations on distribution level introduced by photovoltaic systems , IEEE Transactions on Energy Conversion Volume 21, Issue 1, March 2006, Pages 202-209

[6]. Gansler, W.A. Beckman, S.A. Klein. Investigation of minute solar radiation data. Solar Energy, 55:21-27, 1995.

[7]. Suehrcke H. and Mc Cormikc P.G. Solar radiation utilizability. Solar Energy, 43(6) 339-345, 1989.

[8]. Notton G., Muselli M. and Louche A., Two estimation methods for monthly mean hourly total irradiation on tilted surfaces from monthly mean dayly horizontal irradiation from solar radiation data of Ajaccio. Solar Energy Vol. 57, No. 2, pp. 141-153. 1996

[9]. Boullier P, Le Chapellier M. Colloque Meteorologie et Energies Renouvelables. Valbonne, 1984, p. 632±653.

[10]. Fabero F, Alonso-Abella M, Chenlo F. In: Proc. of the 14th European PV Solar Energy Conference, Barcelone. 1997. p. 2299±302.

[11]. Muselli M, Poggi P, Notton G, Louche A (2000) Classification of typical meteorological days from global irradiation records and comparison between two Mediterranean coastal sites in Corsica Island. Energy Conversion and Management 41: 1043–1063

[12]. Maafi, S. Harrouni, Preliminary results of the fractal classification of daily solar irradiances Solar Energy 75 (2003) 53–61

[13]. Harrouni S., Guessoum A., and Maafi A., Classification of daily solar irradiation by fractional analysis of 10-min-means of solar irradiance Theoretical and Applied Climatology, 80, 27–36 (2005)

[14]. Soule A., Salamatian K., Taft N, Emilion R., Classification of Internet flows. Sigmetrics 2004, New-York

[15]. Emilion R., Process of random distributions. Afrika Stat, Vol. 1, 1, 27-46, 2005.

[16]. Emilion R., Classification et mélanges de processus. C.R. Acad. Sci. Paris, 335, série I, 189-193, 2002