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

Ted Soubdhan, Richard Emilion, Rudy Calif

Abstract

In order to characterize the fluctuating nature of solar radiation in 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 (16.27N, 61W) where important fluctuations can be observed even within a period of a few minutes. The results exhibit four distinct classes of distributions corresponding to different types of days. The sequence of such classes can be of interest for future weather prediction.

Keywords: Clearness index $k_t$; Daily distributions; Dirichlet mixture; Classification; Solar radiation; Solar systems

1. Introduction

In tropical climate, solar radiation is a very fluctuating data, notably due to 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/m$^2$ and occur within a short time interval, from a few seconds to a few minutes depending on the geographical location. These variations fundamentally depend on cloud size, speed and number.

Studies of solar energy systems are traditionally performed using daily or hourly data (Notton et al., 2002; Suehrcke and McCormick, 1988). 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 (Tovar et al., 1998; Woyte et al., 2007).

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 (Woyte et al., 2006). Therefore, in power distribution grids with a high density of PV, rapid fluctuations of the produced electrical power may appear, leading to unpredictable variations of node voltage and power in electric networks. In small grids, such as those that exist on islands (Guadeloupe, FWI), these fluctuations can cause instabilities.

Hence, management of the electrical network and of the alternative power sources requires a better identification of these small time scale 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 surges.

In this paper, we first summarize the variations of a daily solar radiation by a histogram. This histogram is an estimation of the measured daily solar radiation distribu-

Corresponding author.
E-mail address: ted.soubdhan@univ-ag.fr (T. Soubdhan).
tions. These measurements were performed in Guadeloupe with a sampling rate of 1 Hz. We then classify these histograms by estimating a finite mixture of Dirichlet distribution. This yields four classes of distributions corresponding to four types of days. The sequence of these classes can be of interest for future weather prediction.

This paper is organized as follows. Section 2 concerns 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 as well as the main ideas of our method. In Sections 5 and 6, we present the theoretical framework of this method. In Sections 7 and 8 we present the classification results and an analysis of the sequence of classes. In Section 9 we discuss the significance 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, located at 16°15N latitude and 60°30W longitude. In such a tropical zone, solar radiation is important climatic data to be taken into account. The average solar load for a horizontal surface is between 4 kWh/m² and 7 kWh/m² per day. Constant sunshine combined with the thermal inertia of the ocean makes the air temperature variation quite weak, between 17 and 33 °C with an average of 25–26 °C. Relative humidity ranges from 70% to 80% and the wind trades are relatively constant throughout the year.

Our measurements sampled at 1 Hz were performed during one year, from September 2005 to December 2006. Few authors have performed measurements with such a time step (Gansler et al., 1995; Suhrecke and Mc Cormick, 1989; Notton et al., 1996). The measurements were made with a pyranometer from KIPP&ZONEN, model SP-Lite. This sensor was used because it has a response time inferior to 1 s, which is in accordance with our sampling rate. The SP-Lite measures the solar energy received from the entire hemisphere (i.e., 180° field of view) and is equipped with a levelling device and bubble gauge so that it can be perfectly aligned. After one year of on site measurements, the SP-Lite was recalibrated with reference to a Kipp & Zonen CM22 sensor under natural sunlight in clear sky conditions. The reference sensor pyranometer CM22 is stored and used solely for this purpose. After the recalibration test, we found a difference of less than 1% between the two sensors. It should be noted that the manufacturer KIPP&ZONEN recommends recalibrating the SP-Lite after two years of use.

All the data were recorded by a CAMPBELL SCIENTIFIC data logger.

Fig. 1 below, presents two examples of measurements and their corresponding probability density functions (pdf).

3. Data representation

We consider a sample \( (h_1, \ldots, h_n) \) 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 \( h_t \) of \( k_t \) has therefore \( l \) bins, say \( h_t = (h_{t1}, \ldots, h_{tl}) \), such that \( h_{tl} \geq 0 \) and \( \sum_{l=1}^{20} h_{tl} = 1 \).

In the sequel it will be important to consider that conditionally to the observed data, \( h_t = (h_{t1}, \ldots, h_{tl}) \) 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 the obtaining of nonempty bins.

4. Daily solar radiation classification

4.1. An overview of solar radiation classification

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

For example, Boullier and Le Chapellier (1984) 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 (using parameters other than solar radiation such as air temperature, wind speed, humidity...). 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. (1997) end up with nine “typical” day periods by decomposing each studied day into three periods. This study, performed in Madrid (Spain), uses 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. (2000) propose a classification methodology with parameters defined by hourly clearness index profiles. A Ward aggregation classification method is then applied to these parameters. The study done with a monthly (respectively an annual) time step leads to 3, 4 or 5 classes (respectively 3 classes).

Maafi and Harrouni (2003) and Harrouni et al. (2005) use fractal dimension and a daily clearness index \( k_t \) as classification parameters. Three classes obtained by some specific thresholds of these parameters correspond to clear, 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. Such histograms are nonparametric estimators of distributions...
so that they incorporate many parameters of a day, such as all order moments or tail distributions.

Thereafter a classification algorithm is performed on the sample of histograms so that we get classes of daily distributions. The variability of the distributions within a specific class is represented by a random distribution which depends on this class. A well-known example of random distribution is the Dirichlet distribution, suitable for representing a wide variety of distributions. The variability of the distributions within a sample of histograms so that we get classes of daily distributions.

To estimate the parameters of the mixture as well as the parameters of each Dirichlet distribution, we use the Stochastic Approximation Expectation-Maximization (SAEM algorithm) (Celeux and Diebolt, 1992; Emilion, 2002), which is a stochastic version improving the celebrated Expectation Maximization Algorithm (EM algorithm) Dempster et al., 1977; Hartley, 1958; McLachlan and Krishnan, 1997.

The above method is interesting not only for capturing the entire range of daily solar radiation behaviour with all its statistical characteristics, but also for dealing with almost any type of distribution and 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 (Soule and A., 2004). This paper seeks to contribute to the study of 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 \((\Omega, F, P)\) to the space \(P(V)\) of all probability measures defined on a fixed measurable set \(V\). \(\Omega\) represent the space of measurements and \(P = F(\Omega)\). 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 well-known 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 \(\gamma(z_1, 1), \ldots, \gamma(z_l, 1)\) respectively, where \(\gamma(a, b)(x) = \frac{1}{\Gamma(a)} (1 - x)^{b-1} e^{-x} x^{a-1} \int_{[0, +\infty)}(x) dx\).

If we normalize each random variable \(Z_i\) by the sum \(Z = Z_1 + \ldots + Z_l\), then the distribution of the random vector \(X = (\frac{Z}{Z}, \ldots, \frac{Z}{Z})\) is called the Dirichlet distribution \(D(z_1, \ldots, z_l)\). Observe that since \(\frac{Z}{Z} \geq 0\) and \(\sum_{j=1}^{l} x_j = 1\), \(X\) is a random distribution on \(P(V)\), \(V = \{1, \ldots, l\}\). Hence

Please cite this article in press as: Soubdhan, T. et al., Classification of daily solar radiation distributions ... Sol. Energy (2009), doi:10.1016/j.solener.2009.01.010
the Dirichlet distribution is a first natural example of distribution of a random distribution, that is a probability measure on $P(V)$, $V = \{1, \ldots, l\}$.

It can be shown that the $I - 1$ dimensional random vector $(\frac{x_1}{x_0}, \ldots, \frac{x_{I-1}}{x_0})$ has the following density with respect to the Lebesgue measure on the set:

$$T_i = \{x = (x_1, \ldots, x_{I-1}) \text{ such that } x_j \geq 0 \text{ and } \sum_{j=1}^{I-1} x_j = 1\},$$

so-called Dirichlet density:

$$F(x_1, x_2, \ldots, x_{I-1}) = \frac{\Gamma(a_1 + a_2 + \cdots + a_I)}{\Gamma(a_1) \cdots \Gamma(a_I)} x_1^{a_1-1} \cdots x_{I-1}^{a_{I-1}-1}$$

$$\prod_{j=1}^{I-1} (1 - \sum_{j=1}^{I-1} x_j)^{a_j-1}$$

for $(x_1, \ldots, x_{I-1}) \in T_i$.

5.3. Mixture of Dirichlet distributions

Let $X : \Omega = P(V)$ be a RD, with, as above, $V = \{1, \ldots, l\}$ and $P(V) = \{x = (x_1, \ldots, x_l) \text{ such that } x_j \geq 0 \text{ and } \sum_{j=1}^{l} x_j = 1\}$.

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 $D(x_1^k, \ldots, x_l^k)$:

$$P_X = \sum_{k=1}^{K} p_k D(x_1^k, \ldots, x_l^k)$$

with $p_k > 0$ and $\sum_{k=1}^{K} p_k = 1$.

The mixture problem consists of estimating the parameters $p_k$ and $(x_1^k, \ldots, x_l^k)$ $i \in [1, K]$. To achieve this aim, we use an iterative algorithm where the inputs are the $n$ histograms vectors $h_i = (h_{i1}, \ldots, h_{in})$, $i = 1, 2, \ldots, n$ as described before and the number of classes, $K$.

5.4. Dirichlet distribution properties

The Dirichlet distribution has the following property which is particularly useful. If $X = (X_1, \ldots, X_l)$ has a Dirichlet distribution, $D(x_1, \ldots, x_l)$, then the marginal distribution of each component $X_i$ follows a beta distribution:

$$X_i \sim \text{B}(\alpha_i, \beta_i)$$

where $\alpha = \sum_{i=1}^{l} x_i$ is the mass value. Recall that the Beta distribution $\text{B}(\alpha, \beta)$ is defined by the probability density function on $[0; 1]$:

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

These properties make the Dirichlet distribution very attractive for modelling random distributions. Moreover, Dirichlet distributions, and more specifically the mixtures of Dirichlet distributions, are very suitable for encompassing a very large spectrum of various distributions appearing in the real world (Emilion, 2005; Gansler et al., 1995; Suehrcke and Mc Cormike, 1989).

6. Estimation procedure

In the present work we want to classify daily pdfs of $k_t$ based on the similarity of their distributions. We assume that the observed empirical histograms are coming from a source governed by a random distribution. Instead of finding a single distribution to represent all the histograms, 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 $\sum_{k=1}^{K} p_k D(x_1^k, \ldots, x_l^k)$ where component $D(x_1^k, \ldots, x_l^k)$ describes class $k$, and $p_k$ represents 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 independent and identically distributed outcomes of a random distribution $X$ defined by 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 algorithm.

6.1.1. Description of the algorithm

- **Initialization step**: Assign randomly each histogram $h_i$, $i = 1, \ldots, n$ to a class.
- **Simulation step**: Generate randomly $t_i^{(0)}(i = 1, \ldots, n)$ representing the initial a posteriori probability that a histogram $i$ is in a class $k$ where $1 \leq k \leq K$ For $q = 0$ to $Q$ do.
- **Stochastic step**: Generate random multinomial numbers $e_{iq} = (e_{iq}^k)_{k=1, \ldots, K}$ following the probability distribution $(t_i^{(q)})$ where all the $e_{iq}^k$ are 0 except one of them equal to 1. We then get a partition $C = (C_k)_{k=1, \ldots, K}$ of the set of histograms. If $\sum_{i=1, \ldots, N} e_{iq}^k < c(n)$ for some $k$ then return back to the initialization step.
- **Maximisation step**: Estimate the mixing weights:

$$p_k^{(q)} = \frac{1}{n} \left(1 - \gamma_q \right) \sum_{i=1, \ldots, N} t_i^{(q)} + \gamma_q \sum_{i=1, \ldots, N} e_{iq}^k$$

and the parameter value:

$$\lambda_k^{(q)} = \left(1 - \gamma_q \right) \sum_{i=1, \ldots, N} e_{iq}^k \frac{b(f)}{t^{(q)}_i} + \gamma_q \sum_{i=1, \ldots, N} e_{iq}^k \frac{b(f)}{t^{(q)}_i}$$

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

$$t_i^{(q+1)} = \frac{p_k^{(q+1)} D(x_1^k, \ldots, x_l^k)}{\sum_{k=1, \ldots, K} p_k^{(q+1)} D(x_1^k, \ldots, x_l^k)}$$

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 also represents the number of components in the Dirichlet mixture model.

The above algorithm has been tested with $K = 2, 3, 4, 5, \ldots$. 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 (clear sky conditions) and $S = 0$ otherwise.
- Cloudy level ($C$) with $C = 0$ for no clouds, $C = 1$, and $C = 2$ for a completely cloudy day
- 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).

These 3 characteristics were defined after having completed the classification. They are introduced to highlight the dynamic and energy level associated with each class.

7.1. 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 $k_t = 0.75$ (Fig. 2). These distributions are representative of clear sky conditions days of solar radiation with very few clouds and thus a very slow dynamic, as shown in Fig. 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 28 days is 7%.

7.2. 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 value around $k_t = 0.75$ (Fig. 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 Fig. 5. The weight of this class of 126 days is 35%.

7.3. 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 $k_t = 0.1$ (Fig. 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 Fig. 7). In this case the solar radiation is mainly scattered by clouds. The weight of this class of 23 days is 6%.

7.4. 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 values, one around $k_t = 0.25$

and the other one around $k_t=0.75$ (Fig. 8). These distributions are representative of days with significant sunshine combined with a large number of small clouds with high speed of passages and thus with high dynamic levels (Fig. 9). In this class we have 188 days, or 52% of the measured data. This class shows a remarkable bimodality according to previous studies (Maafi and Harrouni, 2003; Soubdhan and Feuillard, 2005).

7.5. Mean pdf and weight of each class

The mean pdf of each class and the associated weight is plotted in Fig. 10 and respectively in Fig. 11.

According to (Fig. 11), class 1 and class 3 days are in the minority as they correspond to rare meteorological events. Class 4 days which are intermittent cloudy, compose the majority 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 Fig. 12. This can represent the yearly evolution of the type of solar radiation days.

This sequence has some interesting statistical properties, such as an exponential residence time distribution in each class (Fig. 13) and a transition from one 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.

9. Interest for solar PV systems

As stated 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 dataset studied, 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 levels, 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 studied site it only represents 28 days over a year.
10. Conclusion

We have first summarized the daily solar radiation fluctuations in tropical climate by the distributions of the clearness index $k_t$, estimated by histograms having $l = 20$ bins. Then, conditionally to the measured 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 Emilion and R (2002), Emilion (2005) 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 ($16^\circ 2^\prime$ N, 61 W) where important fluctuations can be observed even within a period of a few minutes. The results highlight 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 (7% and 6% respectively) while Classes 2 and 4 are large (35% and 52% respectively), but we have chosen to keep Classes 1 and 3 because they have a physical sense: they represent the two extreme conditions that can be encountered, completely clear and completely cloudy.

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.

References


