A Dirichlet-multinomial mixture model-based approach for daily solar radiation classification

Azeddine Frimane, Mohammed Aggour, Badr Ouhammou, Lahoucine Bahmad

A Dirichlet-multinomial mixture model-based approach for daily solar radiation classification

Âzeddine Frimane[a,*], Mohammed Aggour[a], Badr Ouhammou[a], Lahoucine Bahmad[b]

[a] *Laboratory of Renewable Energies and Environment (LR2E), Faculty of Science, IBN TOFAIL University , B.P. 133-14000 Kenitra, Morocco*
[b] *Laboratory of Magnetism and Physics of High Energy (PPR-13), Faculty of Science, MOHAMMED V University , B.P. 1014 Rabat, Morocco*

Abstract

A challenging problem in the classification of daily solar radiation is the selection of the appropriate model complexity and size that best describe the data. This paper introduces a new nonparametric Bayesian method for automatic classification of daily clearness index, by assuming Dirichlet process as a nonparametric prior on the model parameters. Nonparametric methods are free from the parametric model assumptions, and there is no need to specify any parametric specifications, or to restrict the number of classes. Our approach relies on the inference of the posterior distributions using the collapsed Gibbs sampler. The proposed method is tested using measurements from 2003 to 2016, at the Silver Lake monitoring station in the USA (121°3'W, 43°7'N), with a 5-min logging interval. By applying our classification algorithm, three classes of daily clearness index distributions are identified, corresponding to three types of sky cloudiness, namely cloudy, partially cloudy, and clear sky. The proposed classification framework can facilitate the design of solar radiation conversion systems.

Keywords: Solar radiation classification, Bayesian nonparametric, Dirichlet-multinomial, Gibbs sampler.

*Corresponding author
Email address: Azeddine.frimane@uit.ac.ma (Âzeddine Frimane)

1. Introduction

Resource assessment is critical for solar energy—thermal and photovoltaic (PV)—applications. Knowledge of solar radiation at a given location depends on the availability of data. However, ground-based meteorological network is often spatially sparse, and the measurements often span only a short period of time. In addition, even if the measured data is satisfactory, it is often in an unanalyzed form, and the solar system designers cannot utilize it as is. The generalization of the solar energy usage is obstructed by its variability, namely, irregular production over time. This variability depends heavily on weather conditions, and creates problems for the power system operators, whose role is to constantly balance production and energy consumption. There is a large number of innovative ideas and methods to help the solar community in dealing with this problem, including classification strategies (Perez et al., 2016; Lauret et al., 2016; Pérez-Ortiz et al., 2016; Ghonima et al., 2012; Soubdhan et al., 2009).

Classification aims to identify categories, or classes, of data which belong together due to their similarities. The classification algorithms lie at the intersection of artificial intelligence and data science. In recent years, the application of these algorithms in the realm of renewable energy has become increasingly important, which has contributed significantly to the improvement and optimization of renewable energy systems. The reader is referred to the review by Pérez-Ortiz et al. (2016).

Many articles dealing with the classification problem have been published in the literature related to renewable energy, due to its numerous applications in summarizing, learning, and modeling purposes. For instance, Sallis et al. (2011) suggested a classification algorithm for wind gust prediction. In Saitwal et al. (2003), an operational cloud classification system is presented. Neural network (Tapakis and Charalambides, 2013) and fuzzy expert system (Pringle et al., 2014) have also been used for solar data classification. In this paper, we focus on using the nonparametric Bayesian framework as a solution to the classification of daily solar radiation. A nonparametric Bayesian model is a Bayesian model whose parameter space has infinite dimensions. In fact, the persistent difficulties that arise in traditional classification methods are the overfitting effect and the prior selection of the unknown number of classes describing the data, which raise the problem of model selection (Rasmussen, 1999). Therefore, with the infinite Bayesian mixture model based on the Dirichlet process, we do not need to adjust the model complexity or specify the number of classes *a priori*; it is treated as a random variable and will automatically be inferred from the data in a fully Bayesian manner. The Dirichlet process was introduced by Ferguson (1973) and remains a cornerstone of nonparametric Bayesian statistics. Our approach relies on the inference of the posterior distributions using Gibbs sampler, which is efficient and effectively avoids the problems trained by optimization-based methods.

In addition, it is a common practice to combine the sequence of class labels resulting from the classification with a hidden Markov model (HMM). HMM is able to extract the dynamic knowledge of this sequence of solar radiation classes, and thus offers the possibility to predict of future class labels. This approach could be used to generate synthetic solar radiation data, which is considered to be an important design input for simulating the expected performance of solar energy systems (Hontoria et al., 2005; Muselli et al., 2001).

The remaining part of the paper is structured as follows. Section 2 introduces a survey of solar radiation classification and the fundamental concept of the proposed strategy. In Section 3, we present the formulation of the infinite Dirichlet mixture by specifying the prior distributions, the conditional posterior distributions, and we provide the Gibbs-sampling algorithm. Experimental results of classification are discussed in Section 4. An empirical example on sizing a stand-alone PV system is presented in Section 5. Section 6 concludes the paper.

2. Basic concepts

2.1. Context of our study

As far as the methodology is concerned, the classification models of solar radiation conditions can be either deterministic or stochastic (i.e., probabilistic). A deterministic approach designs analytical methods based on physical considerations that study the natural processes occurring in the atmosphere and influencing solar radiation, including astronomical coordinates, atmospheric composition, among others. However, many atmospheric processes are complex in nature, and are yet to be fully understood and expressed in deterministic terms. For this reason, such processes have been treated probabilistically on many occasions.

Stochastic classification of solar radiation can be considered as a problem of time series classification. Liao (2005) surveyed the time series classification methods and reassembles them into three general categories: (1) the raw-data-based classification which performs directly with the raw information, (2) the model-based classification which operates indirectly with models made from the new information, and (3) the feature-based classification which acts indirectly with features extracted from the raw data. A variety of classification methods of solar radiation conditions using different statistic techniques have been proposed by various authors. For example, Muselli et al. (2000) presented a methodology that belongs to the first category to classify the daily global radiation on the basis of hourly

clearness index profiles. This classification used the aggregation Ward's method, allowing determination of 3, 4, or 5 classes with a monthly time step, and 3 classes with an annual time step. Soubdhan et al. (2009) proposed a model-based classification that relies on the use of daily distributions of the clearness index k_t by estimating a finite mixture of Dirichlet distributions. This algorithm was tested with different values of K, the authors found that the most suitable one is $K = 4$, which represents the following 4 different types of solar radiation days: clear sky days, intermittent clear sky days, cloudy sky days, and intermittent cloudy sky days. Fortuna et al. (2016) suggested a strategy that belongs to the class of methods working on features-based methods to classify daily solar radiation patterns into four classes, making use of the fuzzy c-means algorithm.

A common disadvantage of the classification methods described above is that they have investigated the classification of days using inflexible models with the number of parameters set *a priori*. The performance of such methods would depend on whether a suitable parametric form of the model can be identified. In practice, it is very difficult to justify such parametric assumptions of the model, especially due to the insufficient prior information and data of solar radiation, which often lead to misleading results when one or more model assumptions are violated (e.g., over-fitting effect) (Rasmussen, 1999). For instance, it is very difficult to know which distance metric to choose using the aggregation Ward's method. An inappropriate or inconsistent choice would almost surely lead to inappropriate and undesirable results. Furthermore, an outstanding issue for these classification methods is that the unknown number of solar radiation classes should be specified in advance, without any guidance to choose the correct number. This limitation creates the problem of adjusting the model complexity. These methods are sensitive to the choice of the initial parametric models, and there is no guarantee that the number of classes found is the correct number.

To overcome these issues, this paper presents a new algorithm that belongs to the category of model-based classification for labeling the different classes of solar radiation days, based on a tool from nonparametric Bayesian analysis known as the Dirichlet process (DP). It is able to automatically infer the appropriate model complexity and size (i.e., correct number of solar radiation classes), without assuming any hypothesis on model parameters; all the parameters of the model and the number of classes are random variables to be learned from the data in a fully Bayesian manner. In this case, our model is more flexible and requires fewer limitations than parametric models to make good inferences, and particularly, presents a simple computational structure using Gibbs sampler, which avoids the local-minima issue of the optimization-based methods (e.g., expectation-maximization algorithm) used by the parametric techniques (Rasmussen, 1999).

A nonparametric approach implies that the number of classes is open-ended, i.e., it is treated as a random variable K, and its value is able to evolve as new data arrive. Accordingly, by considering that the daily clearness index coded as vectors of integers (called occupation numbers) comes from K different components of multinomial distribution, and by assuming Dirichlet process as a nonparametric prior to the set of parameters, our model will automatically infer all the model parameters and the correct model size. Moreover, the Bayesian nature of such a model makes it capable of improving this study with more physical knowledge through the choice of more informative and accurate prior distributions.

2.2. *Database and pre-processing*

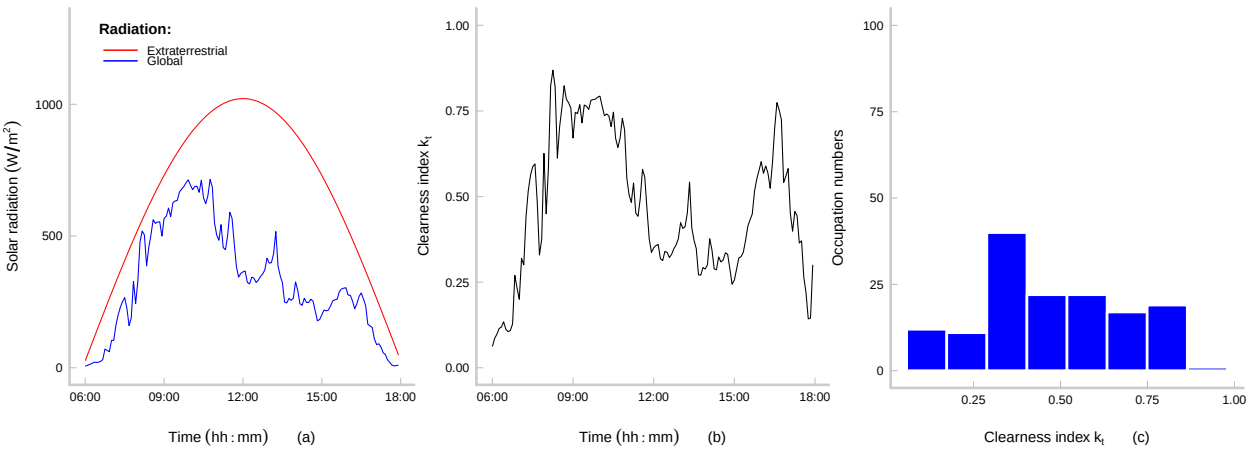

Figure 1: An example of measurements coding: (a) daily solar radiation, (b) corresponding clearness index k_t, and (c) corresponding vector of occupation numbers $\mathbf{X_t}$.

Stochastic solar radiation models are usually performed using high-resolution databases. However, high-resolution long-term solar radiation datasets are rare, mostly due to the high cost of operation and maintenance. In addition, self-

collected data is often incomplete (data gaps), small in size, or not properly quality-controlled, which substantially decreases their usability. Fortunately, in the past few years, scientific community appreciates the interest of making accurate and reliable databases openly available to encourage its reuse for similar or for new purposes. Interested readers can find more detailed information in Yang et al. (2018). Accordingly, the database considered in this work consists of a 5-min time series of solar radiation data collected at the Silver Lake monitoring station in the USA, from 1 January 2003 to 31 December 2016 (without data gaps) and can be freely downloaded from `http://solardat.uoregon.edu/SilverLake.html`.

By definition, the daily clearness index, k_t, is the daily ratio between the global solar radiation H, and the extraterrestrial radiation H_0 on a horizontal surface at a specific location and time. Its value is ranging from 0 to 1. This index is well documented in the literature and is widely used to characterize the solar radiation conditions in different geographical regions. The clearness index is used instead of the clear-sky index because it has a general expression and is calculated without referring to any specific weather conditions of the location (Zhong and Kleissl, 2015). Next, to *code* the clearness index data, the "0–1" interval is first divided the into *class intervals* of equal width. Practically, it is found that the most appropriate number of class intervals is $D = 8$, in line with the Sturges' rule (Scott, 2009). The second step is the introduction of the random D-dimensional vector $\mathbf{X_t}$ with components n_{it} called occupation numbers (ON), which is defined as the count of the clearness index values assigned to each class interval. Fig. 1 illustrates the above-mentioned steps.

Given the geographic coordinates of the Silver Lake monitoring station, the shortest day of the year (the winter solstice, around December 21st) has $\sim$9 daylight hours. However, on the longest day of the year (the summer solstice, around June 21st), the daylight hours are $\sim$15. Consequently, this leads to daily data points between 108 and 180 during the course of the year, considering the 5-min data resolution. Whereas from a methodological point of view, a necessary condition for the ON vectors to be a mixture of multinomial distributions is to be of the same size. Therefore, it is crucial to find an appropriate technique to normalize their sample sizes.

It is important for the time-alignment technique to ensure that the aligned ON vectors preserve all original statistical characteristics, as well as respect the shape of the original ON vectors. Therefore, after choosing the average number of daily data points $n = \sum_{i=1}^{D} n_{it} = 144$ as the optimal size, machine-learning approaches are used to either impute (from short time series to long time series) or resample (from long to short) the time series. In the literature, many data-imputation methods have been proposed, such as dynamic time warping-based imputation (Phan et al., 2017), predictive mean matching (PMM) (Morris et al., 2014), among others. In the present case, the PMM algorithm is used. The PMM algorithm reproduces multiple imputations using Gibbs sampling to best replace the needed data to complete the sample size, and minimize the bias in the imputed values. This iterative method leads to more accurate estimates than a conventional non-iterative imputation technique. Regarding the resampling method, a bootstrap algorithm based on random sampling without replacement is used, as described in Hesterberg (2015). These algorithms are available in many statistical software packages, such as R.

The effect of the alignment operation on the ON vector characteristics can be best observed from the solstices, see Fig. 2 (a) and (c) which represent examples of initial and adapted ON vectors on the winter and summer solstices, respectively. From the summary statistics associated with each vector, it is apparent that this alignment does not lead to a significant change of the ON vectors, neither in shape nor in statistical characteristics for these two extreme cases. That said, the other days of the year are less affected. Fig. 2 (b) shows that the days around the equinox are not affected, since they already have the required size.

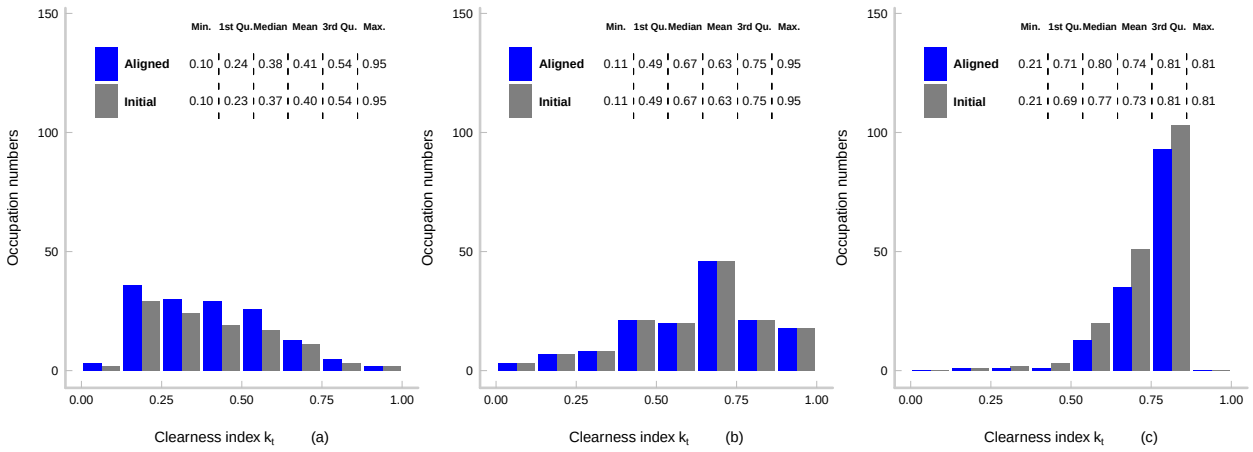

Figure 2: Examples of time-aligned and initial ON vectors with its associated summary statistics: (a) on the winter solstice, (b) on the March equinox and (c) on the summer solstice.

3. Methods

3.1. Bayesian settings

Throughout this section, vector quantities are written in bold. The number of data point N is indexed by t, i.e., $t = 1, \cdots, N$; class interval D is indexed by i, i.e., $i = 1, \cdots, D$; and the component of mixture models K is indexed by j, i.e., $j = 1, \cdots, K$.

The aim of this study is to model the data in terms of infinite Bayesian mixture model to automatically discover interesting patterns in data, and to avoid the overfitting effect in parametric models or traditional approaches based on extensive cross-validation. In general, our learning strategy is to formulate a probabilistic model which suggests that different subsets of the ON vectors $\mathbf{X_t}$ are represented as a mixture of several components. Each component has a simple parametric form (likelihood function) that need to be explicitly specified and justified. In fact, each likelihood function represents an unobserved class of the daily ON vectors, therefore, a specific meteorological regime.

It is known that different sampling rates correspond to different probability distributions of data (Yang et al., 2017). For this reason, a convenient way to specify the distribution from which our ON vectors $\mathbf{X_t}$ are drawn is to initially assume that each day is a collection of clearness index values k_t of size S, and that our data are sampled from this collection with a sample size of $n = 144$. By the fact that solar radiation is a continuous physical function over time, we can then consider S arbitrarily large. Hence, in the limit of large population size S compared with the sample size n, the multinomial distribution can provide an excellent approximation of the density function of the random ON vectors for many realistic sampling schemes (Rice, 2006).

It is important to emphasize that $\mathbf{X_t}$ vectors are not independent variables but conditionally independent given the random parameters of the model. In another way, all the randomness is contained in the random parameters of the model. Following, we considered that each vector $\mathbf{X_t}$ arises conditionally independent from a mixture of multinomial distributions defined by:

$$\begin{aligned}\mathbf{X_t} = (n_{1t}, \ldots, n_{Dt}) \mid n, \{\pi_j\}_{j=1}^{K}, \{\mathbf{p_j}\}_{j=1}^{K} \\ \sim \frac{n!}{\prod_{i=1}^{D} n_{it}!} \sum_{j=1}^{K} \pi_j \prod_{i=1}^{D} p_{ij}^{n_{it}},\end{aligned} \tag{1}$$

where $\pi = \{\pi_j\}_{j=1}^{K}$ is the mixing proportions vector, $\{\mathbf{p_j}\}_{j=1}^{K}$ is the parameter vector for the multinomial components that represent the mean of the class, and $n = \sum_{i=1}^{D} n_{it}$ is the size for the multinomial components.[1]

In Bayesian settings, the choice of prior distribution describes our physical knowledge about how likely we think each class of days will be before any relevant evidence is taken into account. Moreover, this choice can hold or facilitate the posterior computation. Thus, we wanted our priors to have reasonable modeling properties, with an eye to the mathematical and computational convenience through the use of conjugate priors.[2]

[1] The symbol "$\sim$" is read "distributed as" and the symbol "|" is read "conditional to".

[2] If the posterior distributions are in the same family as the prior probability distribution, the prior and the posterior are then called conjugate distributions. Conjugate priors have the convenient feature that, when multiplied by the suitable likelihood, they give a closed-form expression.

DP is a stochastic process whose realizations are probability distributions. It is characterized by two parameters, a positive real scalar known as the concentration parameter α, and the base measure ψ_0 on which the mixture distribution ψ (the random distribution drawn from the DP) over the set of parameters is centered. For computational convenience, a Dirichlet distribution is chosen as the base distribution, with a concentration parameter γ, to achieve the conjugacy to the multinomial distribution:

$$\psi_0(\mathbf{p}) = \frac{\Gamma(\sum_{i=1}^{D}\gamma_i)}{\prod_{i=1}^{D}\Gamma(\gamma_i)}\prod_{i=1}^{D}p_i^{\gamma_i-1}, \tag{2}$$

whereas Γ is the Gamma function.[3]

The random measures ψ can be described explicitly by employing the stick-breaking representation introduced by Sethuraman (1994). We can express this using an indicator variable c_t, which is a positive integer that indicates the latent class associated with the observation t and distributed as π (π is a random probability over the positive integers). Therefore, an equivalent representation of our model is produced by the following conditional distributions:

$$\begin{aligned}
\mathbf{X}_t \mid c_t, n, \{\mathbf{p_j}\}_{j=1}^{K} &\sim \frac{n!}{\prod_{i=1}^{D} n_{it}!}\prod_{i=1}^{D} p_{ic_t}^{n_{it}} \\
c_t \mid \pi &\sim \text{Discrete}\,(\pi_1, ..., \pi_K) \\
\mathbf{p_j} \mid \gamma &\sim \text{Dirichlet}\,(\gamma) \\
\pi \mid \alpha &\sim \text{Dirichlet}\left(\tfrac{\alpha}{K}, ..., \tfrac{\alpha}{K}\right).
\end{aligned} \tag{3}$$

The size of the multinomial components is restricted to be fixed at n. One of the basic ideas of our modeling is that by conditioning on the event $n = 144$, we can compute an equivalent model easily by treating each component i of $\mathbf{X_t}$ independently. This greatly facilitates the computation task. Moreover, all inferences are equivalent whether we consider components as samples from D independent Poisson distributions, or from a single multinomial distribution; this is called the multinomial-Poisson transformation (Baker, 1994). Hence, the multinomial component j can be generated via series of D independent Poisson variables Y_{it} with additional parameters ν_{ij}:

$$Y_{it} = n_{it} \mid \nu_{ij} \sim \exp(-\nu_{ij})\frac{\nu_{ij}}{n_{it}!}. \tag{4}$$

Expressing $\mathbf{X_t}$ with the additional parameters:

$$\begin{aligned}
\mathbf{X_t} &= \left((Y_{1t}, ..., Y_{Dt}) \Big| \sum_{i=1}^{D} Y_{it} = 100\right) \\
&\sim \frac{n!}{\prod_{i=1}^{D} n_{it}!}\prod_{i=1}^{D}\left(\frac{\nu_{ij}}{\sum_{i=1}^{D}\nu_{ij}}\right)^{n_{it}},
\end{aligned} \tag{5}$$

therefore:

$$p_{ij} = \frac{\nu_{ij}}{\sum_{i=1}^{D}\nu_{ij}}. \tag{6}$$

In addition, the joint distribution of D independent normalized gamma[4] distributions is Dirichlet distribution, and each gamma distribution is the conjugate prior to the Poisson likelihood. Therefore, for each ν_{ij} we choose a conjugate gamma prior with the shape parameter γ_{ij} and unit scale:

$$\nu_{ij} \sim \frac{1}{\Gamma(\gamma_{ij})}\nu_{ij}^{\gamma_{ij}-1}\exp(-\nu_{ij}). \tag{7}$$

Likewise, a gamma hyper-prior (shape $= 1$, scale $= 1$) is put on the shape parameter γ_{ij}:

$$\gamma_{ij} \sim \exp(-\gamma_{ij}). \tag{8}$$

Finally, we take a vague[5] inverse gamma prior (shape $= 1$, scale $= 1$) for the concentration parameter α:

$$\alpha \sim \alpha^{-3/2}\exp\left(-\frac{1}{\alpha}\right). \tag{9}$$

[3] The Gamma function Γ is an extension of the factorial function to real and complex numbers.

[4] The gamma distribution is the regularized gamma function.

[5] A vague probability is highly dispersed distribution and it expresses that a wide range of values is plausible.

Algorithm 1 Collapsed Gibbs Sampler

Initialize c_t, for every vector $\mathbf{X}_t$.
for `iter in` $1 : n\_iters$ **do**
  for each j **do**
    for each i independently **do**
      drawing v_{ij} from its posterior conditional on c and hyper-parameter γ_{ij}
      drawing γ_{ij} from its posterior conditional on v_{ij}
    end for
    calculate $\mathbf{p}_j$ by normalizing
  end for
  for t `in` $1 : N$ **do**
    Remove the old assignment c_t for $\mathbf{X}_t$
    if its class being empty **then** remove it and decrease K $(K \leftarrow K-1)$
    end if
    for each j **do**
      Calculate $\mathscr{P}(c_t = j \,|c_{-t}, \mathbf{X}_t, \mathbf{p}_j, \alpha)$
    end for
    Calculate $\mathscr{P}(c_t = K+1 \mid c_{-t}, \mathbf{X}_t, \alpha, \gamma)$
    drawing a new indicator variable c_t from $(\mathscr{P}(c_t = 1), ..., \mathscr{P}(c_t = K+1))$
    assign $\mathbf{X}_t$ to the new class and update K $(K = K \; or \; K = K+1)$
  end for
  update the concentration parameter α .[7]
end for

3.2. Posterior Inference

Our aim is to infer the posterior probability over each indicator variable conditional on all other variables. Subsequently, by integrating over the mixing proportions and letting K tend to infinity, the prior for c_t will have the following limits $\mathscr{P}$ (Neal, 2000):

$$\mathscr{P}(c_t = j; j \text{ already seen} | c_{-t}, \alpha) \longrightarrow \frac{n_{t,j}}{t-1+\alpha} \tag{10a}$$

$$\mathscr{P}(c_t = j \,; \text{new} j | \, c_{-t}, \alpha) \longrightarrow \frac{\alpha}{t-1+\alpha} \tag{10b}$$

where c_{-t} is the collection of all the indicator variables except c_t, and $n_{t,j}$ is the number of the elements in the collection c_{-t} that are equal to j.

Therefore Gibbs sampling for our Dirichlet conjugate mixture model is based on the following conditional posterior probabilities computed by multiplying the multinomial likelihood by the prior:[6]

Probability of a class that currently has data:

$$\begin{aligned}\mathscr{P}(c_t = j \,|c_{-t}, \mathbf{X}_t, n, \mathbf{p}_j, \alpha) \\ \propto \frac{n_{-t,c}}{t-1+\alpha} \frac{n!}{\prod_{i=1}^{D} n_{it}!} \prod_{i=1}^{D} p_{ij}^{n_{it}},\end{aligned} \tag{11}$$

Probability of going to a currently empty class :

$$\begin{aligned}\mathscr{P}(c_t = j \mid c_{-t}, \mathbf{X}_t, \alpha, \gamma) \\ \propto \frac{\alpha}{t-1+\alpha} \frac{n!}{\prod_{i=1}^{D} n_{it}!} \\ \times \frac{\prod_{i=1}^{D} \Gamma(n_{it}+\gamma_i)}{\prod_{i=1}^{D} \Gamma(\gamma_i)} \frac{\Gamma(\sum_{i=1}^{D} \gamma_i)}{\Gamma(\sum_{i=1}^{D}(n_{it}+\gamma_i))}.\end{aligned} \tag{12}$$

3.3. Algorithm

Exact computation of posteriors is intractable, however, we use Markov chain Monte Carlo methods to sample from each posterior distribution (Neal, 2000). We use Gibbs sampler where each variable is sampled in turns from its posterior distribution conditional on all other variables. Computation follows Algorithm 1.

[6] The symbol "$\propto$" is read "proportional to".

[7] The distribution of α is log-concave, so we may draw from this distribution using Adaptive Rejection Sampling (ARS) (Gilks and Wild, 1992).

4. Results

4.1. Computation

Estimating the model parameters requires the computation of the posterior probabilities. 10^6 iterations are performed for modeling purposes (100000 initially for burn-in) to draw independent samples from the posterior distributions given by Eqns. 11 and 12. We select the value of the parameters that maximizes the Bayesian likelihood estimators, sometimes referred to as mode of the posterior distribution.

After running the above algorithm and assigning each vector to the class that gives the highest likelihood, 3 classes are found. These 3 classes correspond to the following three typical day types: (1) clear sky, (2) partially cloudy sky, and (3) cloudy sky.

As suggested by Yang et al. (2018), in order to ensure the readability and reproducibility of the research presented in this paper, and also to encourage other researchers to re-use our algorithm in the same or other applications, we have developed an open-source R package that implements our classification method and shared it in `https://github.com/frimane/SolMultinomClass`.

4.2. Association of classes with the degree of cloudiness

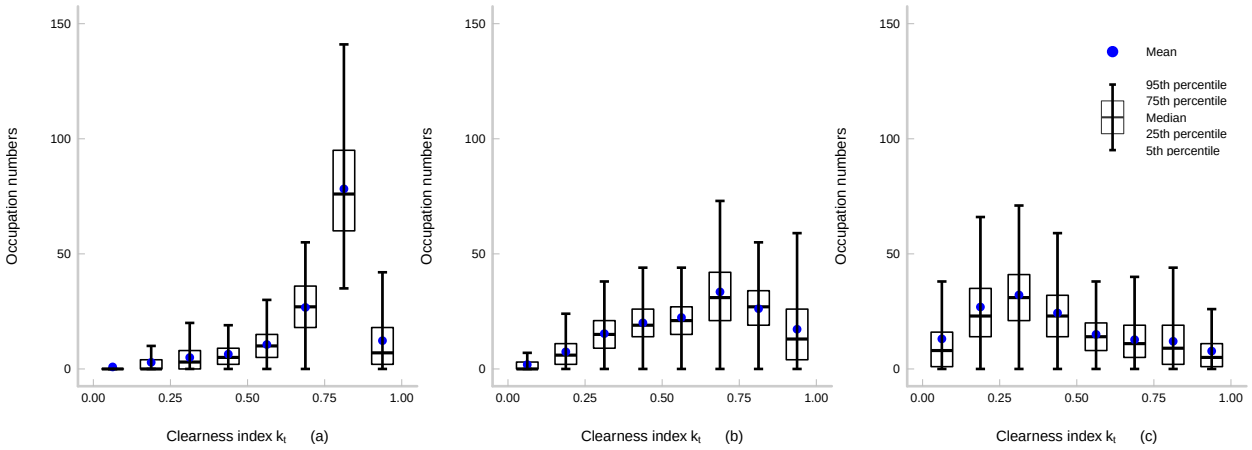

Figure 3: Summary of the classes showing for each class-interval the distribution of the occupation numbers: (a) clear sky, (b) partially cloudy sky and (c) cloudy sky class.

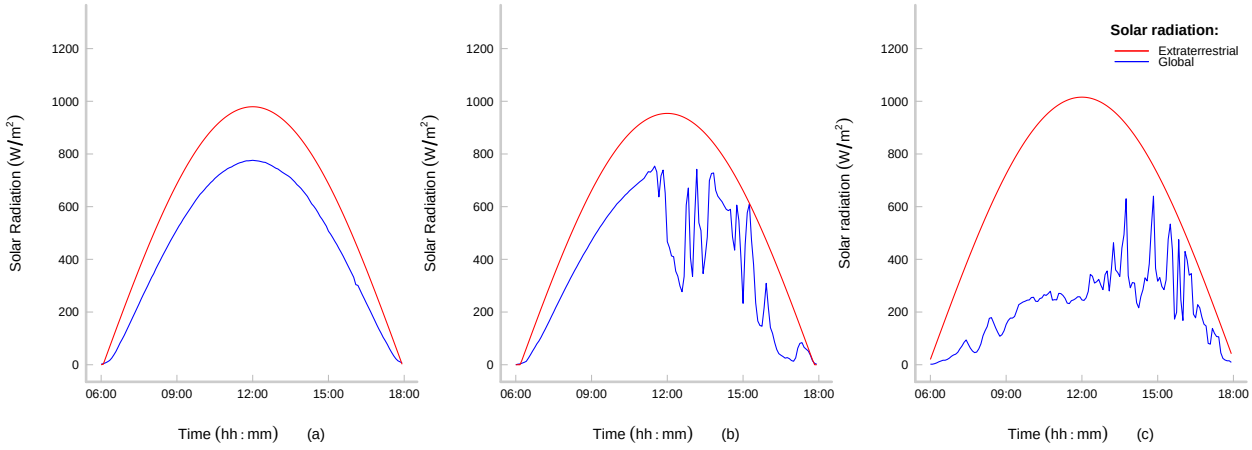

Figure 4: An example of: (a) clear day, (b) partially cloudy day and (c) cloudy day.

The most influential parameters on the propagation of the sun's rays, which reduce the extraterrestrial radiation as it moves through the atmosphere, are atmospheric transparency and cloudiness. Atmospheric transparency depends on the concentration of aerosols, particles, among others. While the cloud effect is due to several aspects, the most important of which are their type and their dynamic.

The clearness index ranges between 0 and 1, and according to their numerical value, we can characterize different weather situations. A low clearness index value implies a small portion of radiation reaching the surface, reflecting an overcast weather situation, whereas a high clearness index value represents a clear-sky condition.

4.2.1. Clear sky days

Under clear-sky conditions, it is evident from Fig. 3 (a) that daily clearness index has a higher occurrence of values around 0.8, when a significant proportion of the global solar radiation reaches the ground. The global solar radiation on clear-sky days is characterized by a cloudless sky or a few high clouds at low speed (e.g., cirrus, cirrocumulus and cirrostratus clouds) assuming a suitable transparency state of the atmosphere (Ohvril et al., 1999). Furthermore, meteorologists use a scale to define cloud cover, and "oktas" is the unit of the scale. Barbaro et al. (1983) determined clear-sky conditions as a function of the cloudiness degree ranging from 0 to 2 oktas. A total of 45% days were classified to this class, the preferred conditions for solar energy conversion systems, especially PV systems, hence deserving greater attention. One representative day from this class is depicted in Fig. 4 (a).

4.2.2. Partially cloudy sky days

Class of partially cloudy conditions is represented in Fig. 3 (b). The variations of the global solar radiation in partially cloudy conditions are greater, and the frequency of occurrence of high-value k_t tends to decrease from that of the clear sky class. This variability of solar radiation and high amplitude of its fluctuations are most likely due to the presence of convective clouds, i.e., cumulus and cumulonimbus (Tomson et al., 2008). Barbaro et al. (1983) limited the degree of cloudiness for partially cloudy conditions between 3 to 5 oktas. A total of 13% of days were assigned to this class. One representative day from this class is shown in Fig. 4 (b).

4.2.3. Cloudy sky days

Class of cloudy days is illustrated in Fig. 3 (c). The proportion of this class is 42%. Typically, in overcast conditions, individual clouds are not present, the reduction of solar radiation quantity is a consequence of multiple opaque clouds location according to the sun's position. It is hypothesized that such situation is caused by clouds at low speed covering most part of the sky, combined with atmospheric turbidity. Barbaro et al. (1983) suggest the limits of cloudy conditions between 6 to 8 oktas. One representative day from this class is shown in Fig. 4 (c).

5. Classes sequence: interest for solar PV systems

The knowledge of solar radiation data from the operating site is a main factor to determine the size of a solar energy conversion system (Rodríguez-Gallegos et al., 2018). Very often, solar systems designers use typical meteorological year data (TMY) to test their designs. However, these modeled TMY data are unable to incorporate all the randomness observed in the actual long-term data. Therefore, synthetically generated data can be useful in solar system design works.

A very common extension to the time series classification is to connect the indicator variables c_t labeling the sequence of the classes into an unobserved Markov chain, resulting in so-called "hidden Markov model" (Muselli et al., 2001). The interest of coupling the classification process with an HMM is to provide a detailed description of the random dynamics of the resulting class sequences. The goal is to have a prediction of the future class labels. Recall that an HMM is a stochastic model in which a sequence of unobserved states are connected via a transition matrix P, and each element in a sequence of observations is generated conditionally on these unobserved states driven by an emission distribution $\mathscr{E}$.

In this section, we investigate the combination of the class sequence obtained using our Dirichlet-multinomial mixture model and an HMM. By approaching the problem in this way, we are able to not only generate daily global radiation sequences capable of reproducing the same long-term statistical characteristics as those measured, but also predict the state of the sky in the short term from the transition matrix. Moreover, such a procedure allows us to summarize 14 years of data in three typical days and a transition matrix only. As a benchmark to analyze the validity of such generator and to allow comparison with results from other methods, an example for sizing a stand-alone PV system in terms of its reliability is presented, using both the measured and generated data.

Fig. 5 shows the fitted transition matrix P underlying the obtained class sequence (C1: clear sky, C2: partially cloudy sky and C3: cloudy sky). The numbers represent the transition probabilities from one class to another. For example, the transition probability from a clear-sky day to another clear-sky day (C1→C1) is 72%, which is interpreted as the residence time in this class. The transition probability from a clear-sky day to a partly cloudy day (C1→C2) is 9%, while the probability of the inverse transition, C2→C1, is 36%. Note that, the emission distribution $\mathscr{E}$ is chosen to follow a multinormal distribution (standard emission distribution) centered by a typical daily clearness index D_s

Algorithm 2 Loss of load probability calculation

```
Setting the site information.
for each value of C_A do
    for each value of C_S do
        for i in 1 : number of days do
            Drawing a state s from P, s ∈ 1,2,3
            Drawing a day D from ℰ(D_s, V_s),
            Simulate the PV system behavior
            Calculate the daily deficit
        end for
        Divide the sum of all deficits by (daily energy consumption of the load * number of days)
    end for
end for
```

defined as the average of the class labeled by the current hidden Markov state s ($s = 1,2,3$). The variance matrix of the emission distribution is chosen to be equal to the empirical variance of the current class V_s.

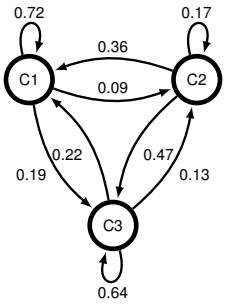

Figure 5: The first order Markov transition matrix underlying the sequence of classes; C1: Clear sky, C2: Partially cloudy sky and C3: Cloudy sky.

It is customary to quantify the reliability of PV systems in terms of the loss of load probability (LOLP). The LOLP represents the confidence level at which the system will satisfy the user load demand, see (Tsalides and Thanailakis, 1986). Several models have been developed for sizing PV systems based on the LOLP approach. In this work, we follow the statistical method presented in Lorenzo (2003). Essentially, this method consists of a detailed simulation of PV system behavior, in order to determine the different combinations of C_A and C_S values that lead to the same LOLP. These value pairs are called the iso-reliability lines. C_A is defined as the ratio between the daily average PV energy production and the daily average load energy consumption. C_S is defined as the maximum energy that can be taken out from the accumulator divided by the daily average load energy consumption. At first, the method involves the introduction of the simulation parameters: C_A and C_S ranges, as well as solar radiation data. The PV array is assumed to have a tilt equals to the latitude of the Silver Lake station, and facing south. For simplicity, the daily energy consumption is assumed constant. It is noted that our method is flexible and offers the possibility of incorporating more complex models of the system elements. A summary of the LOLP calculation using our methodology is shown in Algorithm 2.

Fig. 6 shows a direct mapping between C_A and C_S for three different LOLP values; 0.01, 0.05 and 0.1. The blue curve represents the iso-reliability line obtained from the measured radiation. The red curve represents the iso-reliability line obtained from the generated data. Each point in the $C_A - C_S$ plane represents a size of the PV system. One can choose between a system with a large C_S and a small C_A or vice versa. The optimum is obtained around the point of inflection.

The choice of these values of LOLP is discussed in Lorenzo (2003). In short, for a LOLP < 0.01, PV sizing results are of doubtful quality because of the accuracy limitations associated to the climate variability. Whereas, with a LOLP > 0.1, PV sizing work is inefficient. It should be noted that the methods of LOLP calculation and data generation are also implemented in the same R package mentioned above.

Table 1: MBE(dimensionless), MAE(dimensionless), RMSE(dimensionless) and nMBE(%) of the iso-reliability lines, based on the measured and generated solar radiation data.

	MBE	RMSE	nMBE(%)
LOLP=0.01	- 0.026	0.038	-3.5
LOLP=0.05	0.005	0.008	1.2
LOLP=0.1	0.005	0.007	1.2

Table.1 shows the mean bias error (MBE), root mean squared error (RMSE) and the normalized mean bias error

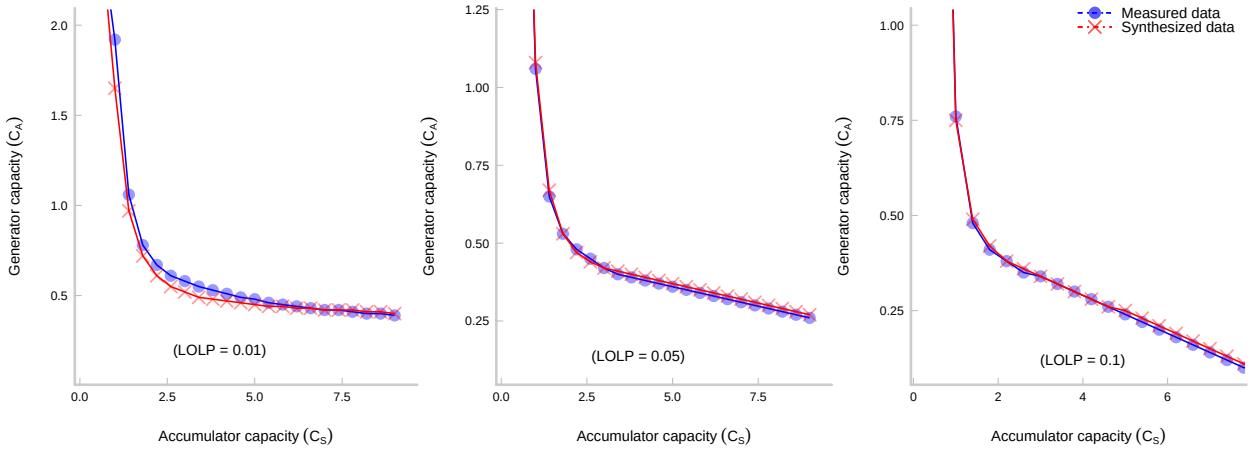

Figure 6: Iso-reliability lines: Generator capacity C_A versus storage capacity C_S based on measured and estimated daily solar radiation for three different LOLP values.

(nMBE) of the iso-reliability lines for each LOLP value. These statistical errors are used to compare the fit of the model generated data to the measured data. As indicated in the table, the RMSE is 0.038 for the reliability level of 0.01 and does not exceed 0.008 for LOLP = 0.05 and 0.1. The MBE does not exceed -0.026 for the different levels of reliability (0.01, 0.05 and 0.1), whereas the nMBE ranges between -3.5% and 1.2%.

By way of comparison, Hontoria et al. (2005) developed an artificial neural network model to generate long-term solar radiation series. In this model, the sizing of PV systems has shown that the RMSE calculated between the real curves and those obtained from simulated data, can reach up to 0.201 for a LOLP of 0.01 and 0.073 for a LOLP of 0.1. Muselli et al. (2001) presented a first order Markov chain model for generating synthetic series of daily global radiation. The sizing of PV systems performed from experimental and simulated data presented a nMBE with a range of -13.3% to 4.2%. These values by far exceed the RMSE and nMBE obtained by our method.

These results prove the efficiency of our method for producing better synthetic data. Our future work will be devoted to the development of a more sophisticated solar radiation sequence generator that only requires the geographical parameters as input (latitude and longitude). This will be accomplished by including additional statistical analysis and incorporating data from several locations, especially that our classification method does not depend on the intrinsic properties of the measurements.

6. Conclusion

In this paper, classification of daily solar radiation as a conjugate nonparametric Bayesian mixture model is studied. A new algorithm is presented. The proposed algorithm is able to automatically determine the correct number of solar radiation classes, and to avoid the overfitting effect of traditional approaches without any adjustment of the model. Firstly, the daily clearness index is summarized as a vector of occupation numbers. Subsequently, the proposed classification algorithm was carried out on these summarized data.

The proposed algorithm was tested using 14 years of 5-min solar radiation measurements at the Silver Lake station, USA. As a result, we have recognized three classes of solar radiation days corresponding to three specific meteorological regimes: clear sky (45%), partially cloudy sky (13%) and cloudy sky (42%).

We have also provided a detailed description of the random dynamic of the resulting sequence of classes, based on the coupling between the Dirichlet-multinomial mixture classification and an HMM. The advantage of this method is that it combines accuracy and simplicity, and can identify the characteristics of solar systems in order to evaluate their performances or their design. knowing that the suggested approach can be generalized to other solar systems and configurations.

This study suggests a potentially fruitful direction for further research in solar radiation classification, especially since it is Bayesian nonparametric. Therefore, it is possible to extend the method of classification with more physical knowledge about solar radiation through the choice of more informative priors or more appropriate non-conjugate priors. In order to facilitate the reproducibility of this work, an R package is implemented and made openly available from `https://github.com/frimane/SolMultinomClass`.

Acknowledgements

The authors would like to thank two anonymous referees who were kind enough to review this manuscript and provided valuable insights and comments. Also, the authors would like to thank the personnel of the Solar Radiation Monitoring Laboratory at the University of Oregon for making the data available.

References

Baker, S.G., 1994. The multinomial-poisson transformation. Journal of the Royal Statistical Society. Series D (The Statistician) 43, 495–504. URL: http://www.jstor.org/stable/2348134.

Barbaro, S., Cannata, G., Coppolino, S., 1983. Monthly reference distribution of daily relative sunshine values. Solar Energy 31, 63–67. URL: http://www.sciencedirect.com/science/article/pii/0038092X83900348, doi:https://doi.org/10.1016/0038-092X(83)90034-8.

Ferguson, T.S., 1973. A bayesian analysis of some nonparametric problems. The Annals of Statistics 1, 209–230. URL: http://www.jstor.org/stable/2958008.

Fortuna, L., Nunnari, G., Nunnari, S., 2016. A new fine-grained classification strategy for solar daily radiation patterns. Pattern Recognition Letters 81, 110–117. URL: http://www.sciencedirect.com/science/article/pii/S016786551630023X, doi:https://doi.org/10.1016/j.patrec.2016.03.019.

Ghonima, M.S., Urquhart, B., Chow, C.W., Shields, J.E., Cazorla, A., Kleissl, J., 2012. A method for cloud detection and opacity classification based on ground based sky imagery. Atmospheric Measurement Techniques 5, 2881–2892. doi:https://doi.org/10.5194/amt-5-2881-2012.

Gilks, W.R., Wild, P., 1992. Adaptive rejection sampling for Gibbs sampling. Journal of the Royal Statistical Society. Series C (Applied Statistics) 41, 337–348. URL: http://www.jstor.org/stable/2347565.

Hesterberg, T.C., 2015. What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. The American Statistician 69, 371–386. doi:https://doi.org/10.1080/00031305.2015.1089789.

Hontoria, L., Aguilera, J., Zufiria, P., 2005. A new approach for sizing stand alone photovoltaic systems based in neural networks. Solar Energy 78, 313–319. URL: http://www.sciencedirect.com/science/article/pii/S0038092X04002397. iSES Solar World Congress 2003.

Lauret, P., Perez, R., Aguiar, L.M., Tapachès, E., Diagne, H.M., David, M., 2016. Characterization of the intraday variability regime of solar irradiation of climatically distinct locations. Solar Energy 125, 99–110. URL: http://www.sciencedirect.com/science/article/pii/S0038092X15006490, doi:https://doi.org/10.1016/j.solener.2015.11.032.

Liao, T.W., 2005. Clustering of time series data–a survey. Pattern Recognition 38, 1857–1874. URL: http://www.sciencedirect.com/science/article/pii/S0031320305001305, doi:https://doi.org/10.1016/j.patcog.2005.01.025.

Lorenzo, E., 2003. Energy collected and delivered by PV modules. Wiley. URL: http://onlinelibrary.wiley.com/book/10.1002/9780470974704, doi:https://doi.org/10.1002/9780470974704.

Morris, T.P., White, I.R., Royston, P., 2014. Tuning multiple imputation by predictive mean matching and local residual draws. BMC Medical Research Methodology 14, 75. doi:https://doi.org/10.1186/1471-2288-14-75.

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. URL: http://www.sciencedirect.com/science/article/pii/S0196890499001399, doi:https://doi.org/10.1016/S0196-8904(99)00139-9.

Muselli, M., Poggi, P., Notton, G., Louche, A., 2001. First order Markov chain model for generating synthetic "typical days" series of global irradiation in order to design photovoltaic stand alone systems. Energy Conversion and Management 42, 675 – 687. URL: http://www.sciencedirect.com/science/article/pii/S019689040000090X, doi:https://doi.org/10.1016/S0196-8904(00)00090-X.

Neal, R.M., 2000. Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics 9, 249–265. URL: https://www.tandfonline.com/doi/abs/10.1080/10618600.2000.10474879, doi:https://doi.org/10.1080/10618600.2000.10474879.

Ohvril, H., Okulov, O., Teral, H., Teral, K., 1999. The atmospheric integral transparency coefficient and the Forbes effect. Solar Energy 66, 305–317. URL: http://www.sciencedirect.com/science/article/pii/S0038092X99000316, doi:https://doi.org/10.1016/S0038-092X(99)00031-6.

Perez, R., David, M., Hoff, T.E., Jamaly, M., Kivalov, S., Kleissl, J., Lauret, P., Perez, M., 2016. Spatial and temporal variability of solar energy. Foundations and Trends in Renewable Energy 1, 1–44. doi:https://doi.org/10.1561/2700000006.

Pérez-Ortiz, M., Jiménez-Fernández, S., Gutiérrez, P.A., Alexandre, E., Hervás-Martínez, C., Salcedo-Sanz, S., 2016. A review of classification problems and algorithms in renewable energy applications. Energies 9. URL: http://www.mdpi.com/1996-1073/9/8/607, doi:https://doi.org/10.3390/en9080607.

Phan, T.T.H., Caillault, Á.P., Lefebvre, A., Bigand, A., 2017. Dynamic time warping-based imputation for univariate time series data. Pattern Recognition Letters URL: http://www.sciencedirect.com/science/article/pii/S0167865517302751.

Pringle, J., Stretch, D.D., Bárdossy, A., 2014. Automated classification of the atmospheric circulation patterns that drive regional wave climates. Natural Hazards and Earth System Sciences 14, 2145–2155. URL: https://www.nat-hazards-earth-syst-sci.net/14/2145/2014/, doi:https://doi.org/10.5194/nhess-14-2145-2014.

Rasmussen, C.E., 1999. The infinite Gaussian mixture model, in: Proceedings of the 12th International Conference on Neural Information Processing Systems, MIT Press, Cambridge, MA, USA. pp. 554–560.

Rice, J.A., 2006. Mathematical Statistics and Data Analysis. Duxbury Advanced. 3 ed., Duxbury Press.

Rodríguez-Gallegos, C.D., Gandhi, O., Yang, D., Alvarez-Alvarado, M.S., Zhang, W., Reindl, T., Panda, S.K., 2018. A siting and sizing optimization approach for PV–battery–diesel hybrid systems. IEEE Transactions on Industry Applications 54, 2637–2645. doi:https://doi.org/10.1109/TIA.2017.2787680.

Saitwal, K., Azimi-Sadjadi, M.R., Reinke, D., 2003. A multichannel temporally adaptive system for continuous cloud classification from satellite imagery. IEEE Transactions on Geoscience and Remote Sensing 41, 1098–1104. URL: http://ieeexplore.ieee.org/document/1206708/, doi:https://doi.org/10.1109/TGRS.2003.813550.

Sallis, P., Claster, W., Hernández, S., 2011. A machine-learning algorithm for wind gust prediction. Computers & Geosciences 37, 1337–1344. URL: http://www.sciencedirect.com/science/article/pii/S0098300411000987, doi:https://doi.org/10.1016/j.cageo.2011.03.004.

Scott, D.W., 2009. Sturges' rule. Wiley Interdisciplinary Reviews: Computational Statistics 1, 303–306.

Sethuraman, J., 1994. A constructive definition of Dirichilet priors. Statistica Sinica 4, 639–650. URL: http://www.jstor.org/stable/24305538.

Soubdhan, T., Emilion, R., Calif, R., 2009. Classification of daily solar radiation distributions using a mixture of Dirichlet distributions. Solar Energy 83, 1056 – 1063. URL: http://www.sciencedirect.com/science/article/pii/S0038092X09000073, doi:https://doi.org/10.1016/j.solener.2009.01.010.

Tapakis, R., Charalambides, A., 2013. Equipment and methodologies for cloud detection and classification: A review. Solar Energy 95, 392 – 430. URL: http://www.sciencedirect.com/science/article/pii/S0038092X12004069, doi:https://doi.org/10.1016/j.solener.2012.11.015.

Tomson, T., Russak, V., Kallis, A., 2008. Dynamic behavior of solar radiation, in: Badescu, V. (Ed.), Modeling Solar Radiation at the Earth's Surface. Springer-Verlag Berlin Heidelberg. chapter 10, pp. 259–281. URL: http://www.springer.com/la/book/9783540774556, doi:https://doi.org/10.1007/978-3-540-77455-6.

Tsalides, P., Thanailakis, A., 1986. Loss-of-load probability and related parameters in optimum computer-aided design of stand-alone photovoltaic systems. Solar Cells 18, 115–127. URL: http://www.sciencedirect.com/science/article/pii/037967878690030X, doi:https://doi.org/10.1016/0379-6787(86)90030-X.

Yang, D., Dong, Z., Lim, L.H.I., Liu, L., 2017. Analyzing big time series data in solar engineering using features and PCA. Solar Energy 153, 317 – 328. URL: http://www.sciencedirect.com/science/article/pii/S0038092X17304796, doi:https://doi.org/10.1016/j.solener.2017.05.072.

Yang, D., Kleissl, J., Gueymard, C.A., Pedro, H.T., Coimbra, C.F., 2018. History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining. Solar Energy URL: http://www.sciencedirect.com/science/article/pii/S0038092X17310022, doi:https://doi.org/10.1016/j.solener.2017.11.023.

Zhong, X., Kleissl, J., 2015. Clear sky irradiances using REST2 and MODIS. Solar Energy 116, 144–164. URL: http://www.sciencedirect.com/science/article/pii/S0038092X15001735, doi:https://doi.org/10.1016/j.solener.2015.03.046.