International Conference on Renewable Energies and Power Quality (ICREPQ'13)
Bilbao (Spain), 20th to 22th March, 2013
Renewable Energy and Power Quality Journal (RE&PQJ)
ISSN 2172-038 X, No.11, March 2013

Clustering daily solar radiation from Reunion Island using data analysis methods

P. Jeanty[a], M. Delsaut[b], L. Trovalet[a], H. Ralambondrainy[b], J.D. Lan-Sun-Luk[a,*], M. Bessafi[a], P. Charton[b], J.P. Chabriat[a]

Université de la Réunion, 15 avenue René Cassin B.P. 7151 97715 Saint-Denis Messag CEDEX 9, île de la Réunion

[a]*LE²P, Laboratoire d'Énergetique d'Électronique et Procédés, Université de la Réunion*
[b]*LIM, Laboratoire d'Informatique et de Mathématiques, Université de la Réunion*

Abstract

A partitioning of daily solar radiation using Hierarchical clustering on Principal Components method is proposed based on daily distribution of direct fraction in solar irradiance. This tool, similar to clearness index, is only based on diffuse and global solar radiation measurement, taking in account the local topology. The clustering method is applied on data collected in Reunion island (21°S, 55°E), a southern subtropical site. We obtain five daily solar radiations classes as results. The interpretations of the average time series signals for each class shows that daily dynamics curves are correlated with local meteorological phenomena. The season distribution of classes is correlated to the intensity of trade winds flow. Thus, it helps us to define an approach to predict the solar radiation for different time horizons and the energy potential of the day.

Key words

Diffuse and global solar radiation; subtropical weather; direct and diffuse fraction; clearness index; Hierarchical Clustering; Principal Components Analysis.

Nomenclature

k_b	direct fraction
k_d	diffuse fraction
f_{diff}	diffuse radiation $W.m^{-2}$
f_{glo}	global radiation $W.m^{-2}$
f_{dir}	direct radiation $W.m^{-2}$
i	day index
n	number of days

1. Introduction

The recent boost in the development of grid-connected photovoltaic power systems lead to new challenges. The character of intermittency, which is the signature of these solar production units, creates a real constraint to electrical grid operators. This problem drastically worsens when the grid concerns a location with limited production facilities and with no possible backup production from outside the place, as it is the case for an isolated island. Besides, the strong meteorological variability which may be encountered on such places strengthens this sensitivity.

The coupling of a power storage system to the production units appears to be a straight forward solution to limit the negative impact of the source intermittency on the grid operation. The design of such additional facility is generally based on the necessary response-time for an efficient backup operation. Finally, size and life-time of such storage plants will directly impact the cost.
The short or medium term prediction of the primary energy source (solar irradiance) is definitively a solution to reduce the storage capacities and, as a result, authorizes to increase the penetration of the photovoltaic units on the power grid.

Therefore, our works consist in defining a prediction model of the sun irradiance. More precisely, this article will present the necessary data treatment prior to prediction modelling. In doing so, we step on the same track as the clustering studies recently performed by [1], [2], [3] and [4]. However, our works follow different patterns by two features. First of all, we introduced k_b, the direct fraction defined as the ratio of the direct component to the global irradiance [5], to replace k_t the clearness index usually found in clustering works. This way, we only deal with local measurements which integrate all environmental influencing factors. The use of k_b also enhances the weight of daily information, eliminating in the results interpretation the competition with the influence of seasons. Besides, k_b is worked out with its time signature allowing us to keep information on the casual link with the meteorological events. It then helps us to interpret the classes from a meteorological point of vue.

In the first part of this document, we will present the location, the instruments used, as well as the measurements main characteristics.

Then a methodology is proposed to cluster time series. It combines proven data-mining methods to identify patterns

[*] Corresponding author.
Email address: J.D. Lan-Sun-Luk <lanson@univ-reunion.fr>; (J.D. Lan-Sun-Luk)

January 21, 2013

https://doi.org/10.24084/repqj11.340

into the data. The approach is applied to different solar schemes on Reunion Island, and the clusters resulted are interpreted taking account meteorological phenomena.

Finally, a first approach on the physical interpretation of the different classes will be presented. Thus, we will highlight the importance of day cycles as a result of our survey on mean day profiles of k_b for each class and their temporal statistical analysis.

2. Data collection and methodologies

A. *Diffuse and global solar radiation measurements*

The solar radiation measurements used for this study were performed in Reunion Island. This Southern Hemisphere volcanic island is exposed to an important solar radiation and is characterized by a humid tropical climate. The combination of a very steep terrain, with large variations in altitude, and prevailing trade winds from south-southeast induces local contrasts in weather patterns at ground level ([6]). Average temperatures oscillate from 25°C to 32°C for coastal regions and from 15°C to 22°C for regions located above an altitude of 1,500 m in the interior of the island ([7]).

The measurements used for this study were sampled at 0.1 Hz, then averaged to give one collected point per minute for final storage. [1], [8] have used the same time step for similar works. Other authors have used a shorter time step, sampling at 1 Hz, as [3], or a longer one, 10 minutes, as [4], [9].

Our measurements were obtained using a SPN1 pyranometer from [10], with a datalogger from Campbell Scientific for data collection and storage. The main point which made us choose the SPN1 sensor is its capability of giving simultaneously the diffuse and the global solar radiation components. Besides, its operation requires just a levelling task when installing, but no regular adjustment (absolutely no mobile part) or other heavy maintenance (except calibration and inner protection against moisture). This compact pyranometer is really an easy-to-use device, its sensitivity and spectral response range being compatible with our goals. This sensor is actually based on a set of seven thermopiles, symmetrically arranged below a shadow dome according to a specific geometry, ensuring that way that, at any time of the day, wherever in the world the measurement is made, there is always one sensor fully exposed to the sun and one sensor fully shadowed. The SPN1 sensor is rated as ”good quality” by World Meteorological Organization.

Raw measurements were collected on Moufia campus location (20° 54′ S, 55° 29′ E) from December 2008 to March 2012, using calibrated instruments leading to a database of 1188 days. For our clustering works, we only retained days with no missing data giving a set of 959 days (80 % of the whole database).

B. *Daily representation of solar radiation*

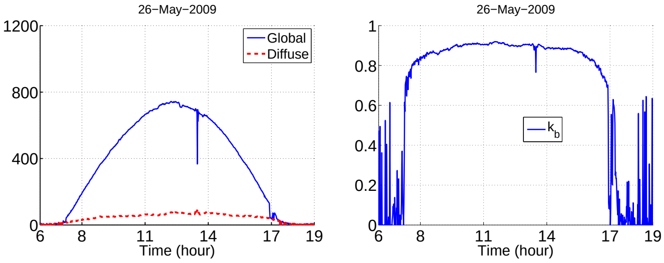
We depict solar irradiance thanks to the introduction of the direct fraction noted k_b, that is related to the usual diffuse fraction k_d as follows:

$$k_b = 1 - k_d = 1 - \frac{f_{diff}}{f_{glo}} = \frac{f_{dir}}{f_{glo}} \tag{1}$$

This way, we define k_b like the ratio of the direct radiation to the global radiation. It is directly obtained thanks to the measurements made at ground level by the SPN1 sensor of f_{glo} and f_{diff} in the conditions described previously.

Thus, when k_b is close to 1, direct radiation level is close to global radiation level, indicating we are in presence of a sunny day. On the other hand, a value close to 0 is the signature of a very cloudy day.

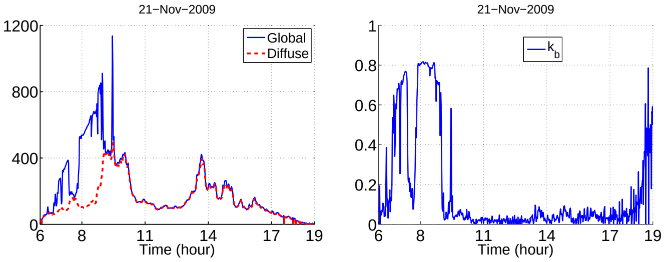
The purpose of our works is now to use the direct fraction k_b to identify different types of days depending on their level of solar irradiance among known situations: clear sky (Fig. 1(a)), cloudy, intermittent cloudy (Fig. 2(a))... To do so, we analysed different samples of daily sequences of k_b measured between 08:00 AM and 05:00 PM (local time, GMT+4). This restricted time slot is chosen to avoid any effect due to solar shade mask actually present at beginning and end of the day.



(a) Solar radiation ($W.m^{-2}$)

(b) Direct fraction k_b

Fig. 1 – Clear sky day example



(a) Solar radiation ($W.m^{-2}$)

(b) Direct fraction k_b

Fig. 2 – Intermittent cloudy day example

C. *Clustering Time Series*

A time serie s_i is a sequence of numbers representing the measurements of a given attribute at equal time intervals. Let $I = \{1, \ldots, n\}$ and $T = \{1, \ldots, p\}$. A set of time series $\{s_i | i \in I\}$ can be represented by a matrix

$$S = (s_i^t)_{i \in I, t \in T} \tag{2}$$

where s_i^t is the value of the attribute for the serie s_i at the time t. In our case, a time serie records the different measurements of k_b every minute during a day. Each serie $s_i = (s_i^t)_{t \in T}$ is viewed as a vector of the euclidian space $E = \mathbb{R}^p$, a weight $p_i = \frac{1}{n}$ is associated to s_i, and the gravity center of the set of series $\mathcal{N}(I) = \{(s_i, p_i)_{i \in I}\}$ is denoted G.

The methodology used to cluster a set of time series consists in combining three data mining methods : Principal Component Analysis (PCA), Ward and K-means clustering methods. We used the package *FactoMineR* ([11]) that implements this strategy in the *R* platform.

1. *Principal Component Analysis* was used as a pre-process for hierarchical clustering method for reduction and denoising. The first principal components are a set of new

uncorrelated variables which extract the main information contained in the data. This issue is important to analyse Time series when measurements are done on a long period, and when the values are correlated. The partitioning obtained by the hierarchical clustering method performed on a set of selected pertinent principal components is more stable than the one obtained from the original data. The number of dimensions suggested by the strategy of *FactoMineR* retains 92% of the total variance.

2. To determine a optimal number of clusters, the *Ward Hierarchical* method ([12]) is applied on these PCA principal components. The Ward method organizes the set of days in a sequence of nested partitions forming a hierarchical tree. The total variance is denoted $V = \Sigma_i p_i d^2(s_i, G)$, and let $P = \{C_k\}_{k\in K}$ a partition of I, p_{C_k} and G_k are respectively the weight and the gravity centre of the cluster C_k. The within-cluster variance $W = \Sigma_{k\in K}\Sigma_{i\in C_k} p_i d^2(s_i, G_k)$, and the between-cluster variance $B = \sum_{C_k\in P} p_{C_k} d^2(G_k, G)$ characterises the homogeneous of the clusters of the partition P. The Huygens theorem allows to decompose the total variance V as follows $V = B + W$. At each step, the Ward method merges two clusters that minimise the reduction of the between-cluster variance. A good number of clusters can be picked out by analysing the between-cluster variance B decreasing of the partitions of the hierarchical tree. This number K of clusters is suggested when the increase of B between $K-1$ and K clusters is much greater than the one between K and $K+1$ clusters.

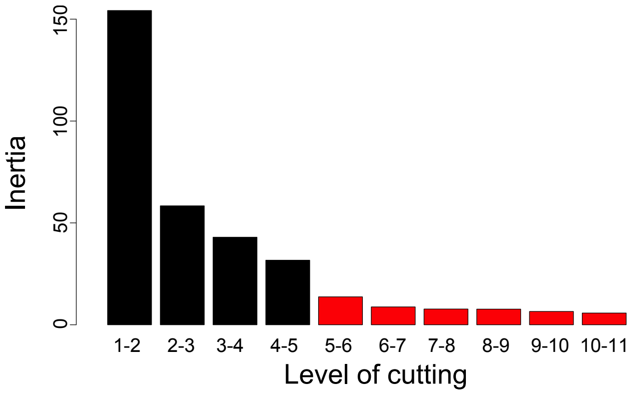


Fig. 3 – Bar plot of the gain in between cluster variance

3. When a number of clusters is selected, the quality of the partition P obtained by cutting the hierarchical tree is improved by applying *K-means* ([13]) algorithm. The quality of the partition P is measured by

$$Q(P) = \frac{B}{V} \qquad (3)$$

the percentage of variance explained by the partition.

D. Cluster interpretation

For a class C_k, the Huygens theorem is written $V_k = B_k + W_k$ with $V_k = \sum_{s_i\in C_k} p_i d^2(s_i - G)$, $B_k = p_{C_k} d^2(G_k - G)$ and $W_k = \sum_{i\in C_k} p_i d^2(s_i, G_k)$. The quality of a class k is measured by

$$Q(C_k) = \frac{B_k}{V_k}\% \qquad (4)$$

the variance explained by the class. A value close to 1 characterizes a homogeneous class.

To give meaning to clusters, we used data-mining tools developed by University of Geneva for extracting interesting knowledge from sequence data. These methods have been successfully applied in various fields: social sciences, bioinformatics... ([14]). These tools are implemented in the *TraMineR* library of *R* platform.

To apply these methods, numerical series must be converted into categorical data. We consider normalized series, each numerical value $s_i^t \in [0,1]$ of a serie s_i is coded in a state or graduation $\hat{s}_i^t$ ranging from 1 to 10 according to his numerical value. Now, a serie $\hat{s}_i$ is a vector of the space $F = [\![1;10]\!]^p$.

Let $\hat{s}^t = (\hat{s}_i^t)_{i\in I}$ the transverse state of the series at the time t. Each cluster will be described by the transverse distribution of states for each time t. The capabilities of *TraMineR* allow us to plot the state distribution for each class (Fig. 4).

E. Classification Results from Réunion Island

In this subsection, we present the results obtained by application of the clustering strategy to Réunion island data. Fig. 3 suggests a partition of the data into 5 classes with a quality of 66%. The table 1 gives the description of each class. Except the class 3, we can see that the weight of each class is similar. We notice that the variance explained is important for the extremes class (1 and 5). It means that these class are the more homogeneous and the least one is class 3.

Class	Number of observation	Freq.	Variance explained
1	178	18.6 %	79.9 %
2	191	20.0 %	59.2 %
3	130	13.6 %	39.5 %
4	223	23.2 %	47.6 %
5	237	24.7 %	78.2 %

Table 1 – Cluster description

Fig. 4 gives the distribution of levels k_b on all days and allows the determination of the significance of each class. Moreover, this representation allows us to look through all days of each class and evaluate the dispersion within the classes. We notice the homogeneity of each class and low frequencies of average levels in graduation for classes 1, 2, 4 and 5. In class 3, the presence of all levels of graduations of k_b suggests a strong variability in days.

3. Physical interpretation of the classes

The five classes partition of the direct fraction resume the sunshine regimes in Reunion Island regardless of seasons, leading to a clustering of the diurnal sunshine between 08:00 AM and 05:00 PM (local time) by average trend of class, as represented on Fig. 5. This clustering shows two types of curves :

- T_1 type for which we observe a very overcast sky at beginning of the day giving anyway relatively low values of k_b for the whole day (Class 1 and Class 3); For the days of T_1 category, two different situations may be identified : one giving a continued slow degradation during the day (Class 1) and another characterized by a quick and frank improvement for the whole day (Class 3);

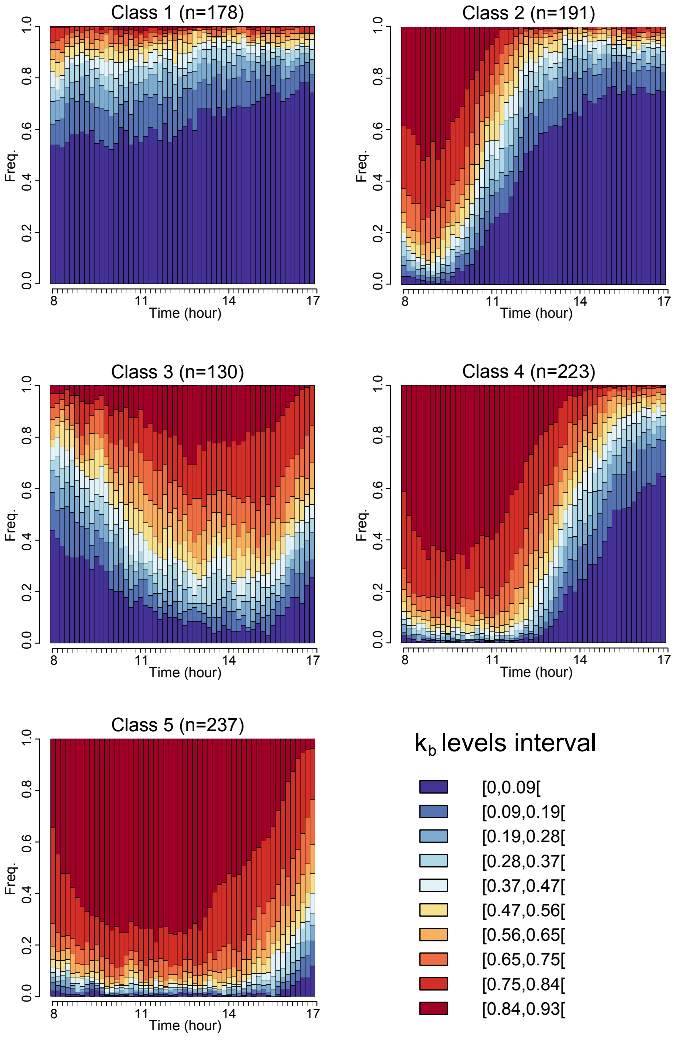


Fig. 4 – Transverse distribution of all states of k_b

- T_2 type that correspond an uncluttered sky giving precedence to high values for k_b (Class 2, Class 4 and Class 5). It may be noted that the T_2 category is relative to days, which start with a nice weather but showing a degradation initiated at different times later in the day. The atmospheric tide probably plays an important role in the variability of the semi-diurnal k_b value noticed for Classes 2 and 4. The different times when the degradation occurs may suppose the superposition of another forcing.

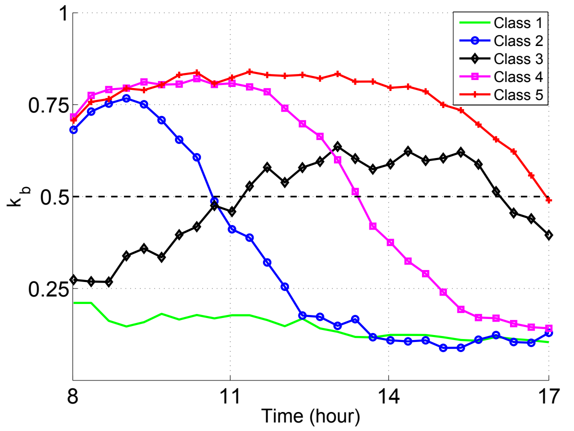


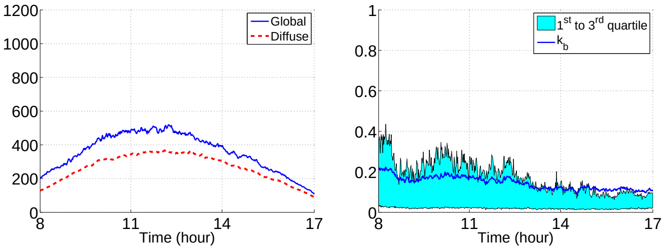
Fig. 5 – Average trend of direct fraction for each class

A. Daily interpretation of classes

To illustrate the main profile of each class, we decided to represent the mean value, the first and the third quartile of the direct fraction, the mean value of the diffuse and global radiations (see Fig. 6, 7, 8, 9 and 10).

A.1. Class 1 : cloudy days

Class 1 corresponds to a very low level of sunshine all day. The consequently averaged value of k_b as seen in Fig. 6(b), indicates a significant cloud cover. The diffuse and global radiations have almost the same shape and are relatively close to each other (Fig. 6(a)). This class presents dominant local phenomena which include, on one hand, the weak trade winds accompanied by a flow of moisture leading to significant effects of orographic clouds, and, on the other hand, the land breeze phenomenon induced by thermal contrasts.

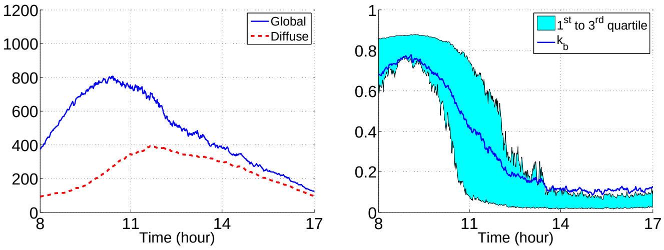


(a) Solar radiation ($W.m^{-2}$) (b) Direct fraction k_b

Fig. 6 – Class 1, mean of class

A.2. Class 2 : intermittent bad days

Class 2 has a sunny beginning until mid-morning around 09:00 - 09:30 AM and a cloudy afternoon (Fig. 7(b)). Diffuse radiation is dominant in the afternoon while the direct component is more important in the morning (Fig. 7(a)).



(a) Solar radiation ($W.m^{-2}$) (b) Direct fraction k_b

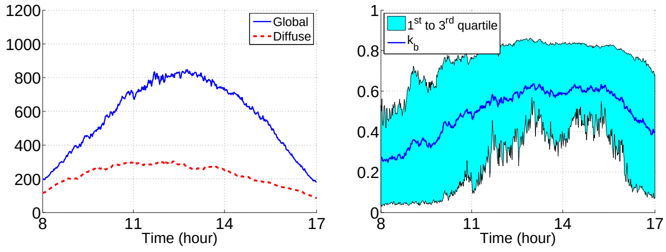
Fig. 7 – Class 2, mean of class

A.3. Class 3 : disturbed days

Class 3 corresponds to a day with a variable weather with a high variability of the direct fraction: improvement in the late morning and moderate cloud cover in the afternoon. In Fig. 8(b), we observe the maximum of k_b close to 0.6. The global radiation illustrates the disturbed weather feature day of this class (Fig. 8(a)).

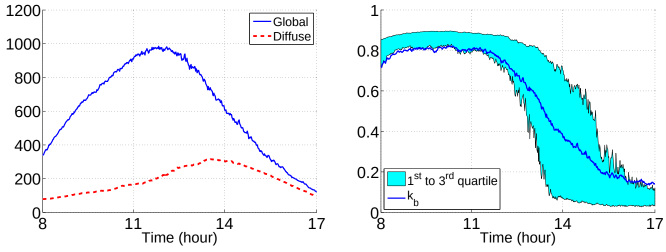
A.4. Class 4 : intermittent good days

The behavior of class 4 is similar to that of Class 2, but with a stronger sunny regime during all morning till early afternoon (Fig. 9). Diffuse radiation takes place later in Class 4 than in Class 2.



(a) Solar radiation ($W.m^{-2}$) (b) Direct fraction k_b

Fig. 8 – Class 3, mean of class

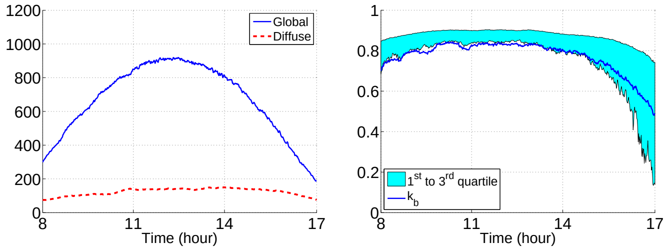


(a) Solar radiation ($W.m^{-2}$) (b) Direct fraction k_b

Fig. 9 – Class 4, mean of class

A.5. *Class 5 : clear sky days*

Class 5 days, characterized by a k_b with very little variations between the first and third quartile (Fig. 10(b)), correspond to a regime of good weather throughout the day. Cloudy overflows affecting the site do not have a systematic character since direct radiation dominates in this class (Fig. 10(a)).



(a) Solar radiation ($W.m^{-2}$) (b) Direct fraction k_b

Fig. 10 – Class 5, mean of class

B. *Seasonal distribution of classes*

The representation of classes over time (Fig. 11) shows the distribution of classes over the years. Note the missing days represented by the shaded areas.

Moreover, we can hypothesize the predominance of Class 5 in austral winter from May to October and Class 2 in austral summer from November to April [7].

In order to analyse of seasonal distribution of classes, we represent Fig. 12 the cumul number of occurences for each month of the year. Class 2 is dominant over the other classes from November to January. This could be explained by the effect of the land breeze blowing from 08:00 AM till 05:00 PM combined with trade winds that are slighter but carrying potential of cloud induced by the orography. Class 5 is dominant over much of the year from March to October with a significant primacy from May to July. Easterly flow is present almost throughout the year but with varying intensity depending on the months of the year. On average, the trade winds are strong during the austral winter (May to October) and lighter during the austral summer (November to April). It reflects the variability of the latitudinal positioning of the Mascarene high. The preponderance of a system of clear sky during the day during the months from May to July suggests that the intensity of trade winds flow is dominant. This consequently limits the local effect of cloud overflow induced by the land breeze that would be observed on the site of Moufia.

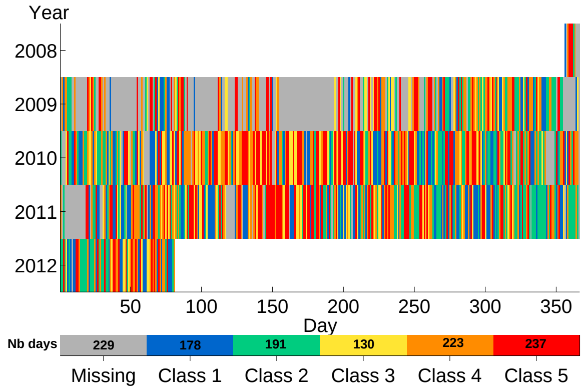


Fig. 11 – Time representation

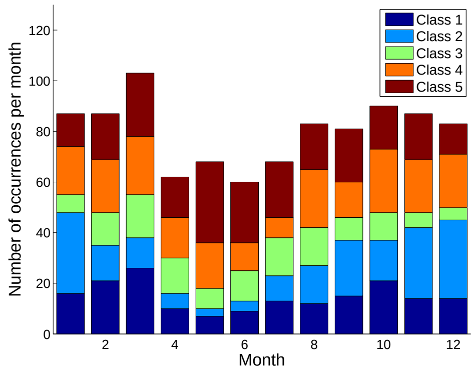


Fig. 12 – Histogram of the annual distribution of the classes

In addition, the flow of austral winter trade winds is more stable and at a lower relative humidity rate compared to the one of the austral summer which bears instability.

It may be noted that these two classes have a very large variability all along the year. However, a difference in the percentage of classes occurrence may be noticed, since we have, on average, around 17 % for Class 1 and 26 % for Class 5.

This implies that diurnal variability is modulated by seasonal variability, although maintaining the five classes.

C. *Classes for solar radiation prediction*

The classification into five classes highlights the intrinsic characteristics of different classes distributed into two types (T_1, T_2). It provides a better understanding of the day and allows to characterize them in the interests of prediction. The contribution of the type, class and their characteristics can be the start point of the implementation tool for the prediction of the solar radiation within a day.

As a matter of fact, the solar radiation between 10 AM to 2 PM (highest energy slot time) could be predicted from information from the beginning of the day. The observation of k_b in the early morning (8 AM - 10 AM) gives us the type of curved line and therefore an information on the class of the day.

The class can be determined by taking into account other physical variables (humidity, pressure, wind ...), slope and level of k_b in the morning, or by calculating the distances to the paragons of each class for the start of the day.

The knowledge of an estimation of the class will finally provide confidence intervals on the value of k_b between 10 AM and 2 PM. For this, we need only to consider the inter quartile of each class which gives us maximum and minimum bounds for the k_b level on the afternoon. (Fig. 6(b), 7(b), 8(b),9(b), 10(b))

This process can be repeated during the morning and day to adjust the choice of the class. In addition, we may take into account the seasonality of subclasses to define and refine the confidence intervals, or subdivide each class by a new classification to quantify the levels of k_b in each class.

4. Conclusion

This paper deals with the problem of mining daily solar radiation data in Reunion Island. Clustering real-world data is not a trivial task, natural or well separated classes are rarely present in such data. We proposed a methodology to cluster time series which combines tree data-mining methods : Principal Component Analysis, Ward and K-means clustering. Applied to our data, interesting patterns related to clusters have been founded.

We obtained five classes which characterize this subtropical weather. The prediction of the next-day classes will be now our priority. Therefore, we shall conduct a cross-correlation analysis between the direct fraction and other meteorological variables.

The clustering of temporal signals from the experimental direct fraction allowed us to highlight the various diurnal cycles in an island environment. The weather conditions in the early morning and mid morning give us relevant information on the class of the day. The establishment of a predictive model will be based on the identification of precursors of these two morning states.

Finally, the possibility of interpreting these classes together with the weather patterns shows that the problem of solar radiation prediction is somehow equivalent to the implementation of a short term meteorological model. This enable us to extend this method to other subtropical region.

Acknowledgement: this work received financial support from Europe, Regional Reunion Island Council and the French government through the ERDF (European Regional Development Fund) and ADEME (French Environment and Energy Management Agency).

References

[1] M. Muselli, P. Poggi, G. Notton, A. Louche, Classification of typical meteorological days from global irradiation records and comparison between two mediterranean coastal sites in corsica island, Energy Conversion and Management 41 (Issue 10) (2000) 1043–1063.

[2] S. Harrouni, A. Guessoum, A. Maafi, Classification of dailysolar irradiation by fractional analysis of 10-min-means of solar irradiance, Theor. Appl. Climatol. 80 (2005) 27–36.

[3] T. Soubdhan, R. Emilion, R. Calif, Classification of daily solar radiation distributions using a mixture of dirichlet distributions, Solar Energy 83 (Issue 7) (2009) 1056–1063.

[4] M. Gastón-Romeo, T. Leon, F. Mallor, L. Ramírez-Santigosa, A morphological clustering method for daily solar radiation curves, Solar Energy 85 (Issue 9) (2011) 1824–1836.

[5] J. Ruiz-Arias, H. Alsamamra, J. Tovar-Pescador, D. Pozo-Vázquez, Proposal of a regressive model for the hourly diffuse solar radiation under all sky conditions, Energy Conversion and Management 51 (2010) 881–893.

[6] G. Taupin, M. Bessafi, S. Baldy, J. Bremaud, Tropospheric ozone above the southwestern indian ocean is strongly linked to dynamical conditions prevailing in the tropics, Journal of Geophysical Research 104, 7 (1999) 8057–8066.

[7] G. Jumeaux, H. Quetelard, D. Roy, Atlas climatique de la Réunion, Météo-France, ISBN 978-2-11-128623-8, 2011.

[8] G. Notton, C. Paoli, S. Vasileva, M. L. Nivet, J. Canaletti, C. Cristofari, Estimation of hourly global solar irradiation on tilted planes from horizontal one using artificial neural networks, Energy 39 (Issue 1) (2012) 166–179.

[9] A. Maafi, S. Harrouni, Preliminary results of the fractal classification of daily solar irradiances, Solar Energy 75 (Issue 1) (2003) 53–61.

[10] Delta-T-Devices, Spn1 - sunshine pyranometer (May 2012).
URL `http://www.delta-t.co.uk`

[11] S. Lê, J. Josse, F. Husson, Factominer: An r package for multivariate analysis, Journal of Statistical Software 25 (1) (2008) 1–18.
URL `http://www.jstatsoft.org/v25/i01`

[12] F. Murtagh, P. Legendre, Ward's hierarchical clustering method: Clustering criterion and agglomerative algorithm, ArXiv e-printsarXiv:1111.6285.
URL `http://adsabs.harvard.edu/abs/2011arXiv1111.6285M`

[13] F. Husson, J. Josse, J. Pagès, Principal component methods - hierarchical clustering - partitional clustering: why would we need to choose for visualizing data, Technical Report Agrocampus.
URL `http://www.agrocampus-ouest.fr/math/`

[14] A. Gabadinho, G. Ritschard, N. Müller, M. Studer, Analyzing and visualizing state sequences in r with traminer, Journal of Statistical Software 40, 4 (2011) 1–37.
URL `http://www.jstatsoft.org/v40/i04/`

[15] R. E. Frank, W. F. Massy, D. G. Morrison, Bias in multiple discriminant analysis, Journal of Marketing Research 2 (1965) 250–258.