Expert Systems with Applications 40 (2013) 7141–7150

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Data mining and statistical techniques for characterizing the performance of thin-film photovoltaic modules

Rafael Moreno Sáez [a,1], Mariano Sidrach-de-Cardona [a,1], Llanos Mora-López [b,2,*]

[a] Departamento de Física Aplicada II, E.U. Politécnica, Universidad de Málaga, Campus de Teatinos, 29071 Malaga, Spain
[b] Departamento de Lenguajes y Ciencias de la Computación, ETSI Informática, Universidad de Málaga, Campus de Teatinos, 29071 Malaga, Spain

ARTICLE INFO

Keywords:
Solar spectral distribution
Photovoltaic modules
Performance ratio
Data mining
Statistical models

ABSTRACT

A method for characterizing the performance ratio of thin-film photovoltaic modules based on the use of data mining and statistical techniques is developed. In general, this parameter changes when modules are working in outdoor conditions depending on irradiance, temperature, air mass and solar spectral irradiance distribution. The problem is that it is usually difficult to know how to include solar spectral irradiance information when estimating the performance of photovoltaic modules. We propose five different solar spectral irradiance distributions that summarize all the different distributions observed in Malaga. Using the probability distribution functions of these curves and a statistical test, we first checked when two spectral distributions measured can be considered to have the same contribution of energy per wavelength. Hence, using this test and the *k*-means data mining technique, all the measured spectra, more than two hundred and fifty thousand, are clustered in only five different groups. All the spectra in each cluster can be considered as equal and the *k*-means technique estimates one centroid for each cluster that corresponds to the cumulative probability distribution function that is the most similar to the rest of the samples in the cluster. The results obtained proves that 99.98% of the functions can be considered equal to the centroid of its cluster. With these five types of functions, we have explained the changes in the performance ratio measured for thin-film photovoltaic modules of different technologies.

1. Introduction

The increase in the performance of solar photovoltaic modules is a pressing issue in order to improve solar photovoltaic market share and make this industry more competitive. There are many elements where this efficiency can be achieved. The elements to be studied to better establish module performance are the key parameters given by the standard test conditions (STC) used to characterize and correct modules as defined by some international standards such as IEC 60904-3 (IEC, 1989) or IEC 61215 (IEC, 1993). The typical standard variables are temperature (25°), irradiance (1000 W/m^2) and AM 1.5 spectrum. This paper is focused on the study and characterization of the spectrum (solar irradiance spectral distribution) and its influence on thin-film photovoltaic devices. The performance of these devices has a greater dependence on the spectrum than photovoltaic modules built using conventional technologies such as mono and policristalin-Si.

Two standards for the AM 1.5 spectrum were defined for terrestrial use by the photovoltaic industry and the ASTM (American Society for Testing and Materials) (ASTM, 2012). The importance of this parameter lies in the fact that photovoltaic modules have a spectral response, (Martín and Ruiz, 1999), which will only allow a part of the whole solar spectral irradiance received on their surface 'to be seen'. Spectral response is a characteristic of every cell technology. Depending on this factor a module will be able to provide more or less energy for one quantity of irradiation received. To study how solar spectral irradiance affects different cell technologies and to explain the different module performances observed, it is first necessary to have an indepth knowledge of the solar spectrum in any outdoor conditions where photovoltaic modules will be operating.

Previous papers used different parameters to characterize solar spectrum. These include the Spectral Factor (SF), (Fabero and Chenlo, 1991), Mismatch Factor (M, MM or MMF), (Poissant et al., 2003), Usefull Fraction (UF), (Gottschalg et al., 2003), Photovoltaically Active Fractions (PAFs), (Berman et al., 1999), and the Average Photon Energy (APE), (Gottschalg et al., 2003). One of the reasons why solar spectrum is not among the typical variables is that measuring solar spectral irradiance involves using very expensive devices. We propose to analyze the APE as this parameter uses a single number to characterize a whole

* Corresponding author. Tel.: +34 951952299.
E-mail addresses: rmorenosaez@ctima.uma.es (R. Moreno Sáez), msidrach@ctimauma.es (M. Sidrach-de-Cardona), llanos@lcc.uma.es, llanos@uma.es (L. Mora-López)rmorenosaez@ctima.uma.es (R. Moreno Sáez), msidrach@ctima.uma.es (M. Sidrach-de-Cardona), llanos@lcc.uma.es, llanos@uma.es (L. Mora-López).
[1] Tel.: +34 951952299.
[2] Tel.: +34 952132802.

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.06.059

spectrum what makes easier for it to be included and treated as another meteorological variable, once calculated. The uniqueness of the relationship between the solar spectral irradiance distribution and the APE is proven, (Minemoto et al., 2009); however, these authors have proven this correspondence by dividing the values of spectral irradiance into several 50 nm bands and using these integrated spectral irradiance values. We propose a different approach, based on the use of a statistical test that enables us to use all the values of the spectrum. Moreover, the authors suggest that the results obtained should be verified with data from locations with different climatic conditions.

In this study we are going to characterize the spectra at the location of Malaga, Andalusia (Spain) with a Mediterranean climate whose typical features are high humidity and mild and warm temperatures. Solar spectral irradiances have therefore been recorded at the laboratory facilities of the Solar Photovoltaic Energy group at the University of Malaga during a period of over one year. By using more than one year of data a complete range of spectrum under real working conditions is obtained and a possible seasonal effect is isolated. Once the spectrum is defined a more precise prediction of the energy produced by modules can be performed. This will result in better tools for the photovoltaic industry to size up photovoltaic plants and their integration into the power grid.

One of the objectives of this paper is to determine the different types of solar spectral irradiance distribution curves and how these different curves are related to the different estimated APE values. Once these relationships are established, further goal of this paper is to analyze how the performance ratio (PR) for different photovoltaic modules is explained using the relationship found. PR values for these modules are estimated using data collected outdoors for over one year.

We propose to use both statistical and data mining techniques that allows us to handle a large quantity of data in an automatic way to analyse and characterize the different APE values in our laboratory. On the one hand, we have used the well known Kolmogorov–Smirnov two sample test to check whether or not two distributions are the same (homogeneity test). On the other hand, we have used the *k*-means data mining technique to cluster all the estimated APEs in several clusters where all the measurements included have the same cumulative probability distribution of solar spectral irradiance.

Many different areas have used clustering including text mining, statistical learning and pattern recognition, (Duda et al., 2001; Hastie et al., 2001; Jain et al., 1999). In previous papers, spectra were gathered using the APE value but data mining techniques will here be used to characterize the spectra in a few types, aimed at trying to put into the same group spectra with different values of APE but with the same relative distribution of the irradiance for each wavelength.

Once the different solar spectral irradiance measurements are characterized taking into account the similarities observed, this information has been used to analyze the performance ratio of different modules for many different outdoor conditions.

In the second section, a detailed description of the data set is presented. In the third section the values of performance ratio for thin-film modules of different technologies are depicted and analyzed depending on several parameters. The fourth section is dedicated to analyzing the relationship between solar spectral irradiance and APE value using statistical techniques and to characterizing the different types of spectra irradiance measured using the *k*-means data mining technique. The fifth section is dedicated to presenting the obtained results and discussing these results for their use in the characterization of performance ratio of thin-film photovoltaic modules. Finally, the last section summarizes the conclusions of the work.

2. Data set

The device used is a grating spectroradiometer that works on the spectral range of visual and near-infrared where most of the cell technologies have their spectral response. The model used for this study is a Grating Spectroradiometer EKO MS-710 prepared for continuous outdoor exposure. It has a silicon sensor that allows a spectral measurement ranges from 330 to 1050 nm, VIS and NIR. The collection time is reduced to between 10 ms and 5 s, so that a fine spectrum with no distortion is obtained on days with moving clouds. This device is placed on a fixed 21° slope frame where continuous spectra were captured during a period of over one year, in the Solar Photovoltaic Energy group facilities at the University of Málaga (Spain), latitude 36.7°N, longitude 4.5°W, height 50 m. The spectroradiometer has a glass dome. To avoid dust and other substances being deposited on the dome, daily cleaning and maintenance are carried out in order to measure solar spectral irradiance as more accurate as possible. There are some other photovoltaic devices close to the spectroradiometer such as piranometers, calibrated cells and modules of different technologies on the same tilted frame structure and placed on the same plane, with the same angle of inclination as recommended in the international standard IEC 60904-7 (IEC, 1995).The manufacturer provides its own software to use the spectroradiometer to measure solar spectra. A custom-built system for I–V curves developed by Pilougine et al. (2011), at the group of PV systems of the University of Malaga was used to collect the meteorological data and the electrical module parameters.

For over a full year, solar spectral irradiances were collected all day long from sunrise to sunset at the rate of one spectrum per minute between November 2010 and May 2012, both included. Solar spectral samples were measured under the full range of meteorological conditions expected in Malaga. This meant that snow was never recorded, but heavy rain, humidity, wind, cloud intervals and sun were experienced. Based on these premises, around four-hundred-thousand spectra were captured with a spectral resolution below 8 nm at wavelength interval of 0.75 nm.

The data recovered for the module were obtained with a time interval of five minutes where their electrical parameters I_{SC}, V_{OC}, T_{MOD}, G and I–V curves plus meteorological data are captured simultaneously.

Although these measurements are very accurate the exact time when sun rises and sets is very difficult to implement in the measurement system, and therefore some of the spectra were captured in the dark at night. Therefore, the data was filtered to avoid reflection produced by angle of incidence and other effects. Spectra taken with an elevation angle under fifteen degrees are removed, following (Nann and Riordan, 1991).

Eq. (1) was used (Reda and Andreas, 2004) to calculate the topocentric elevation angle without atmospheric refraction correction -α- for a surface oriented in any direction.

$$\alpha = \arcsin(\sin\varphi \cdot \sin\delta' + \cos\varphi \cdot \cos\delta' \cdot \cos H') \quad (1)$$

where:

φ is the geocentric latitude
δ' is the topocentric sun declination
H′ is the local hour angle for the sun transit

The total amount of data remaining after filtering is up to 70% of the initial measurements, a total amount of two-hundred-eighty-thousands spectra.

Some thin-film modules and the above described spectroradiometer are installed nearby in the same frame. Three commercial thin-film modules are used for this paper: an a-Si/μc-Si, a single

a-Si and a CdTe with peak power quoted by the manufacturer of 121 W_P, 60 W_P, 70 W_P, respectively.

3. Performance of thin-film photovoltaic modules

The influence of solar spectral irradiance distribution on photovoltaic modules is the final scope of this study. After obtaining a complete characterization of this distribution for different APE values, the relationship between APE, temperature of the module and photovoltaic module performance ratio (PR) is achieved. The energy produced by photovoltaic modules and their subsequent performance is related to the temperature, air mass and solar radiation. Moreover, some authors also include the influence of the Average Photon Energy (APE), (Zanesco and Krenzinger, 1993), that contains information on the solar spectral irradiance distribution. In our paper, we are going to follow indications given by Minemoto et al. (2007) to connect spectrum by means of APE and energy produced where it is necessary to calculate the performance ratio (PR) of the module and depict it as a function of APE and T_{MOD}, using contour graphs. The performance ratio applied to a stand-alone module is the final yield divided by the reference yield (Kymakis et al., 2009). This final yield is defined as the energy output of the photovoltaic module divided by its peak power and the reference yield is the in-plane solar irradiance divided by the array reference irradiance 1000 W/m^2, according to the following expression:

$$PR = \frac{P_{output}/I_{input}}{P_{STC}/1000} \tag{2}$$

where:

P_{output} is the module output power, (W), I_{input} is the received irradiance in the module surface, (W/m^2), P_{STC} is the module power at standard conditions, (W) and 1000 is the irradiance at standard conditions, (W/m^2).

The performance ratio is measured in % and represents the deviation of the output of the module with respect to the standard conditions (STC): if the PR value is 100, the module is working at STC and output power is its nominal output power. Different PR values represent different working conditions of the module due to different irradiance, air mass, module temperature and/or incident spectrum.

The aim of studying PR as a function of APE, solar spectral irradiance and temperature is to establish the most typical spectrum and temperature conditions for the module working in Malaga and how these affect the performance of the module. If the relationship between solar spectral irradiance distribution and APE is established, it would allow the PR of thin-film modules of different technologies to be better explained. These parameters are used to perform a complete characterization of the usual conditions, temperature-irradiance-air mass-spectrum. Spectrum is inherent to APE and irradiation to PR.

We have analyzed the dependence of the performance of the installed thin-film modules on the temperature and APE. The results can be seen in Fig. 1 where the PR is depicted as a function of APE and temperature. As it can be observed, the operation conditions of modules are very different from STC conditions where PR is 100.

As it can be observed from these figures, the different PR values estimated cannot be explained only using the APE and temperature values. For instance, for the temperature of 25, there are several different values of PR and APE that correspond to the same value of temperature. Greater knowledge of the spectral distribution that corresponds to the different APE values would allow us explain the different PR values estimated.

4. Characterization of solar spectral irradiance distribution

Average Photon Energy, APE, is an indicator of the light that reaches the surface of the Earth. To calculate APE (Williams et al., 2003) the integrated irradiance has to be divided by the integrated photon flux density (Minemoto et al., 2007), according to the following expression.

$$APE = \frac{\int_a^b E(\lambda)d\lambda}{q\int_a^b \Phi(\lambda)d\lambda}\ [eV] \tag{3}$$

Spectral characteristics of received irradiance change depending on the particles suspended in the atmosphere, daytime or season, and therefore its corresponding APE also changes. There is also high dependency on the location mainly due to latitude and atmospheric agents. These can be grouped in two principal effects, Rayleigh Scattering and Water Vapor, (Ishii et al., 2011).

The value of APE for AM 1.5 in the range between 350 and 1050 nm is 1.88 eV. Lower value of APE denote a shift of the spectrum to the 'red' while a higher APE value means a higher contribution of the smallest wavelengths that is the same as a shift to the 'blue' (Nann and Emery, 1992). Hence, the APE value changes depending on air mass and the atmospheric composition as the solar spectral irradiance distribution changes.

We have estimated the APE value for all the spectra we have recorded. The histogram of the distribution of estimated APEs is shown in Fig. 2(A) and (B) shows the cumulative histogram.

It can be seen from Fig. 2(B) that over 70% of the spectra have APE of over 1.88 eV (APE value for the STC spectrum AM 1.5), which means that the spectra are shifted to the blue. This can be explained because of the water vapor held in the atmosphere at Malaga, a seaside town with high rates of relative humidity. Water vapor filters spectral irradiance in higher wavelengths so the contribution of this part of the spectrum decreases and a shift to the blue occurs (Ishii et al., 2011).

4.1. Comparing distributions depending on the APE value

In order to determine the relationship between APE and solar spectral distribution, a comparison of spectra with the same APE is performed using all the recorded measurements. The goal of this comparison is to evaluate the uniqueness of this relationship for the different values of APE, as proposed by Minemoto et al. (2009) using data for Tokyo. Nevertheless, the method we propose to use is different as it is based on the whole data measured for the spectrum and on the use of a statistical test. We have checked the uniqueness of this relationship for many different values of APE ranging from 1.79 eV to 2.09 eV that correspond to different values of air mass and meteorological conditions.

For instance, two spectral irradiances can be seen in Fig. 3. Both were recorded on different days and moments, one at midday and the other in the evening. These spectra are different at first glance, but they have the same APE value.

Although the spectra of Fig. 3 are different, the proportional contribution of irradiance for each wavelength to the total energy is the same. Some normalization of the spectrum values is necessary to obtain these contributions. Instead of normalizing by irradiance as in Minemoto et al. (2007), we propose to normalize using the total irradiance for the 350 to 1050 nm range. When this normalization is used, the data obtained are the cumulative distribution probability functions (c.p.d.f.) that range from 0 to 1. Mathematically, this function is estimated using the following expression:

$$c.p.d.f.(\lambda) = P(\Lambda \leqslant \lambda) = \int_{-\infty}^{\lambda} f(t)dt \quad \text{for } -\infty < \lambda < \infty \tag{4}$$

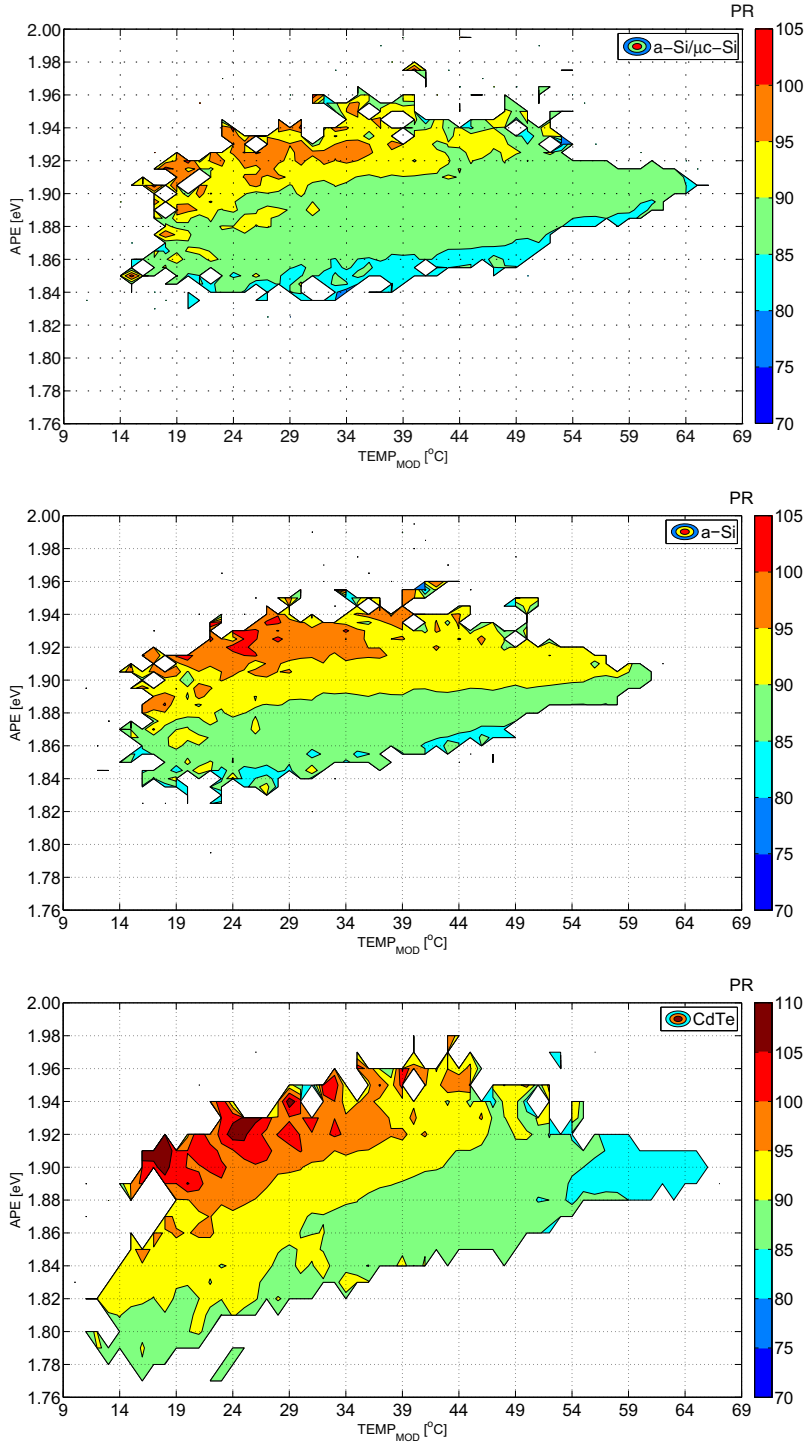

Fig. 1. Performance ratio vs average photon energy and temperature.

When expression (4) is used for the irradiances in Fig. 3 the c.p.d.f. shown in Fig. 4 are obtained. Both c.p.d.f. curves have the same slope and shape, which means that spectra do have the same proportional contribution of irradiance in all the wavelengths. To reinforce this explanation a superposition of both c.p.d.f. curves can be seen in Fig. 5.

The overlapping is almost total. Consequently both spectra have the same distribution (proportion) of irradiance in every wavelength. It could therefore be said that these spectra are equal in relative terms.

In order to compare all spectra that have the same APE value (± 0.05) we have used the well known Kolmogorov–Smirnov two sample test that allows us to determine whether or not the distributions of two variables are the same. This test can be performed using the Kolmogorov–Smirnov statistic that compares the empirical c.p.d.f.s obtained with each sample; this statistic for two samples (1 and 2) of sizes m and n, is estimated using the expression:

$$dist_{m,n} = \max_{t\in\mathbb{R}} |\hat{cpdf}_1(\lambda_t) - \hat{cpdf}_2(\lambda_t)| \quad 350 < \lambda_t < 1050 \tag{5}$$

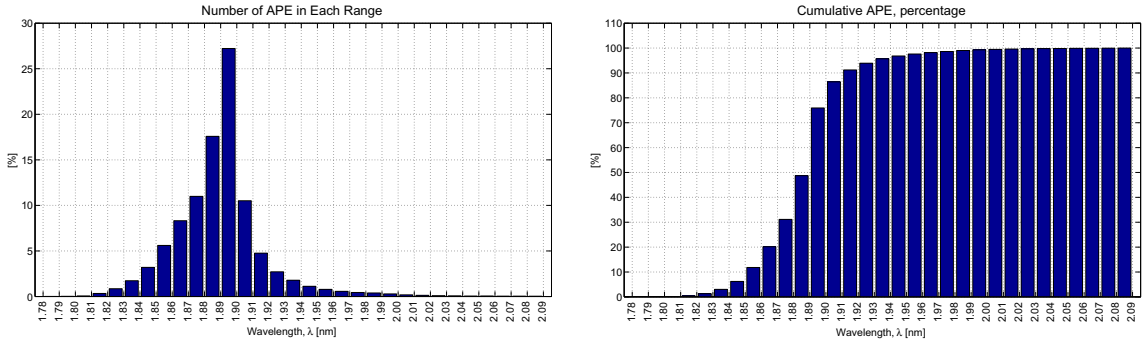

Fig. 2. Histogram (A) and cumulative (B) probability distribution of calculated APE from spectra measured.

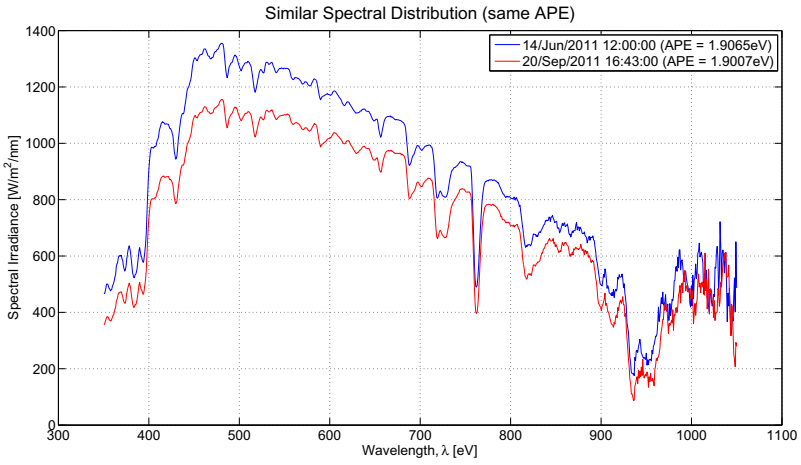

Fig. 3. Two spectra with the same value of APE.

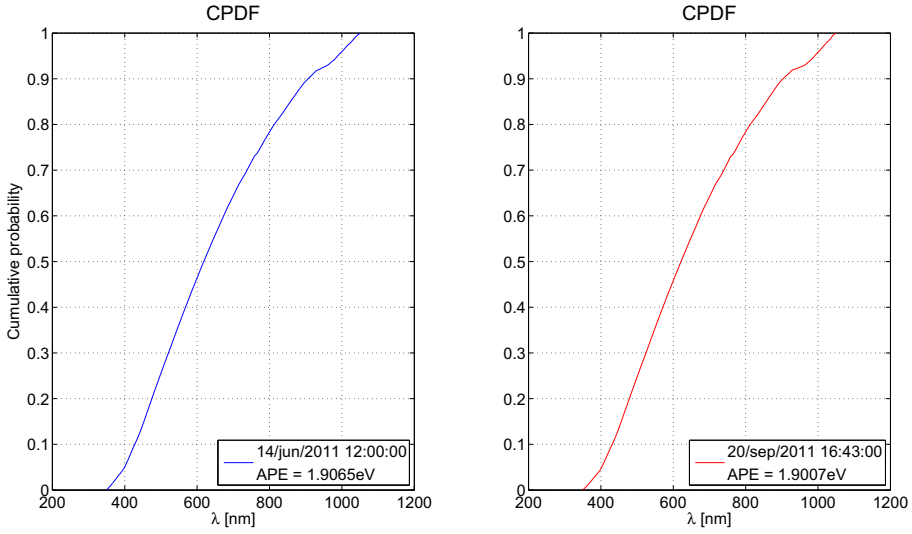

Fig. 4. Cumulative probability distribution function for two spectra that have different spectra but the same APE.

where:

$|A|$ means the absolute value of A. The distribution will be the same with significance level α if the following holds:

$$\left(\frac{nm}{n+m}\right)^{1/2} dist_{n,m} < c_\alpha \tag{6}$$

where c_α is a critical value that only depends on α (for details, see e.g. Rohatgi and Saleh, 2001).

To carry out spectrum comparison they are gathered in APE classes using width intervals of 0.01 eV. Solar spectral irradiances with the same APE value are then compared by means of the c.p.d.f. (4). The intervals used are the following:

$$X_i \in G_i \text{ if } APE_i \in [1.78 + (i-1) * 0.01, 1.78 + i * 0.01) \tag{7}$$

To demonstrate that this procedure detects spectra which wavelengths contributes differently to the total APE, a comparison of two spectra that are different (captured in different days and

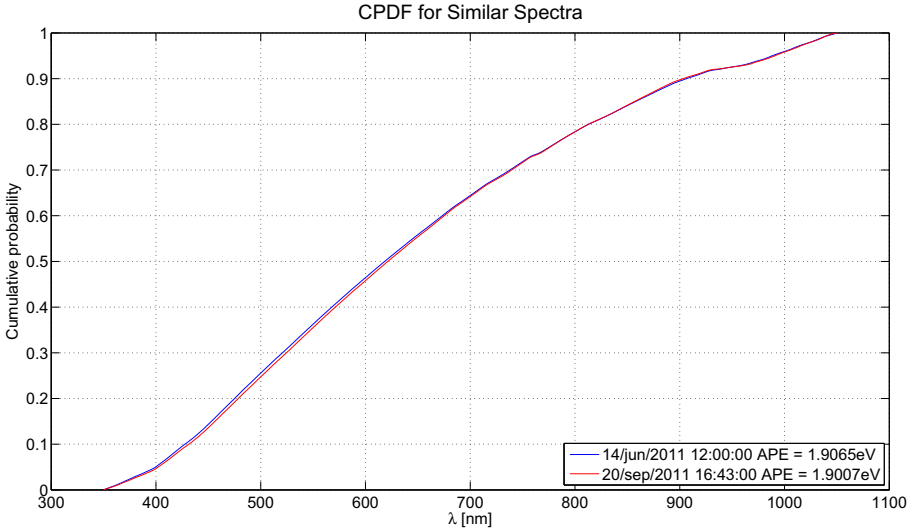

Fig. 5. Cumulative probability distribution function overlapped for the two spectra of Fig. 3. These measurements have different spectra but the same APE.

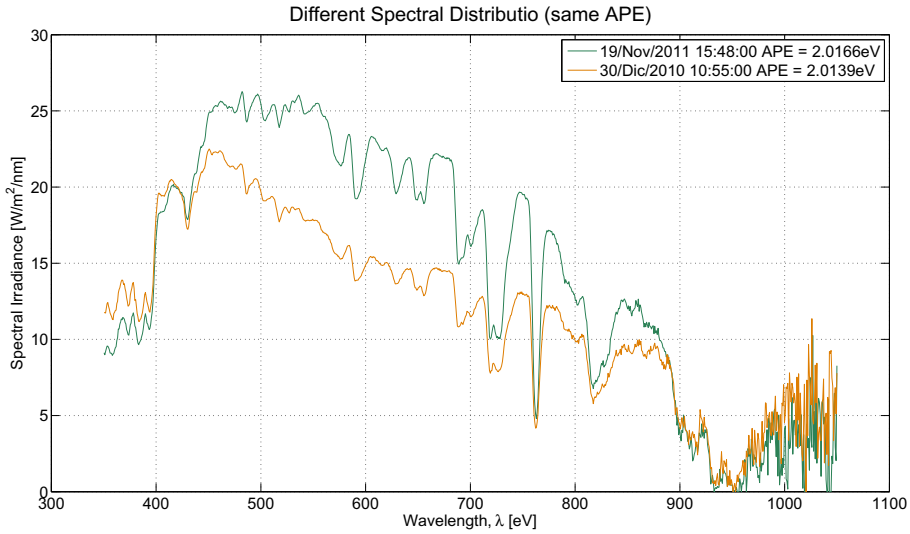

Fig. 6. Two different spectra that have the same APE value but different contribution for each wavelength.

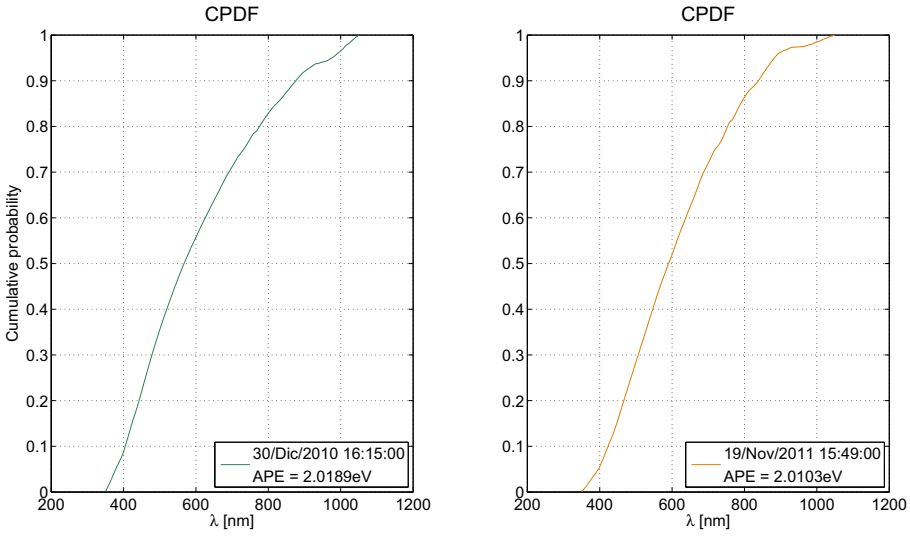

Fig. 7. Cumulative probability distribution function for two different spectra with the same APE.

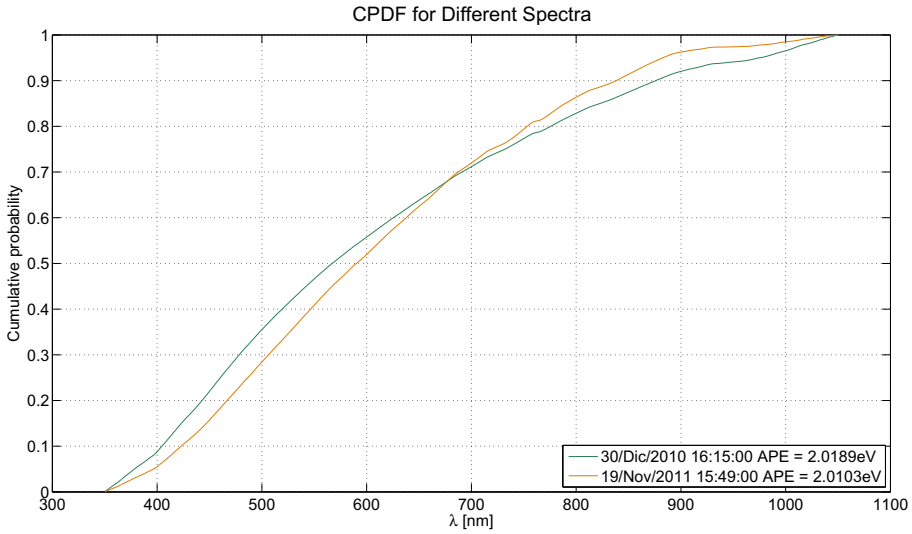

Fig. 8. Different cumulative probability distribution functions overlapped for two different spectra with the same APE.

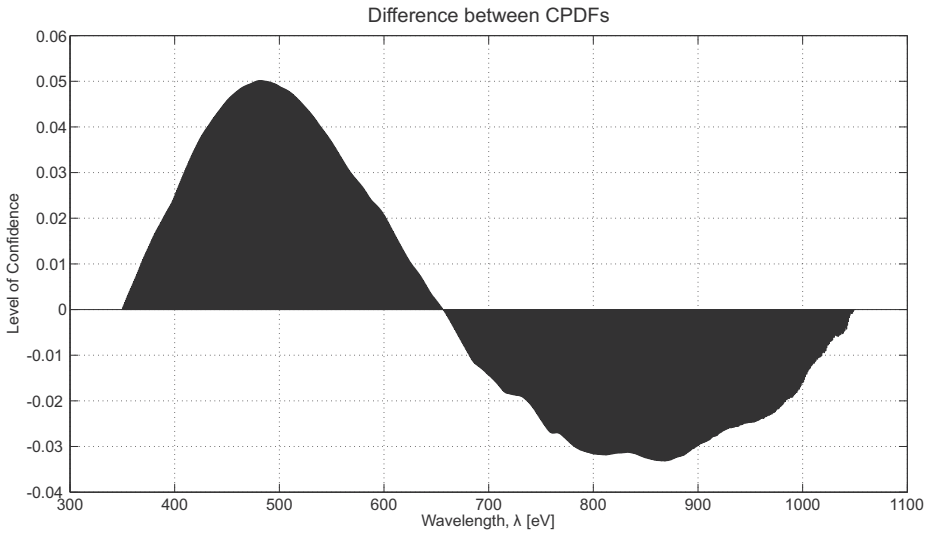

Fig. 9. Distance of two different cumulative probability distribution functions that are statistically different, using the Kolmogorov–Smirnov two samples test.

moments) and which have the same APE value are carried out, see Fig. 6. The result of calculating the c.p.d.f. of both spectra can be seen in Fig. 7. An overlapping of both c.p.d.f.s is performed in Fig. 8 where it can be seen that these spectra have a different contribution of irradiance per wavelength. The distance between c.p.d.f.s is calculated in Fig. 9.

These spectra are statistically different as the maximum distance between these two curves is greater than the statistic c_α for α as significance level.

4.2. Types of solar spectral irradiance distribution

We likewise analyse whether the equality occurs among spectra with different APE. For this analysis, we first cluster the spectra. The cluster method chosen is the *k*-means method (MacQueen, 1967). It is an unsupervised learning algorithms that calculates for each cluster the Euclidean distance between gathered samples and reference samples named centroids and assigns each sample to the closest centroid iteratively, (Han and Kamber, 2001). The procedure is quite easy and has been implemented several times. The goal of clustering is to partition the whole data set into a number of groups (Cao and Liang, 2011) so that data in the same cluster are more similar than data in other clusters. Then, for each cluster, the algorithm selects the data (in our experiment the c.p.d.f.) that is more similar for all the spectra in the cluster, that is, the centroid of the cluster.

The steps of clustering using *k*-means are the following

1. Select k centroids, one per each cluster, as far as possible each other,
2. Assign every sample to a cluster (centroid),
3. Recalculate the centroids taking into account the distance among samples in each cluster and its centroid,
4. Repeat steps 2 and 3 until the centroids no longer move.

Clustering the spectra allows us to compare all the c.p.d.f. of one cluster and its centroid using the Kolmogorov–Smirnov test. The difficult decision in our case is to choose the number of clusters. The maximum number of clusters is given by the number of APE classes, thirty one. After several trials five clusters were settled as optimal.

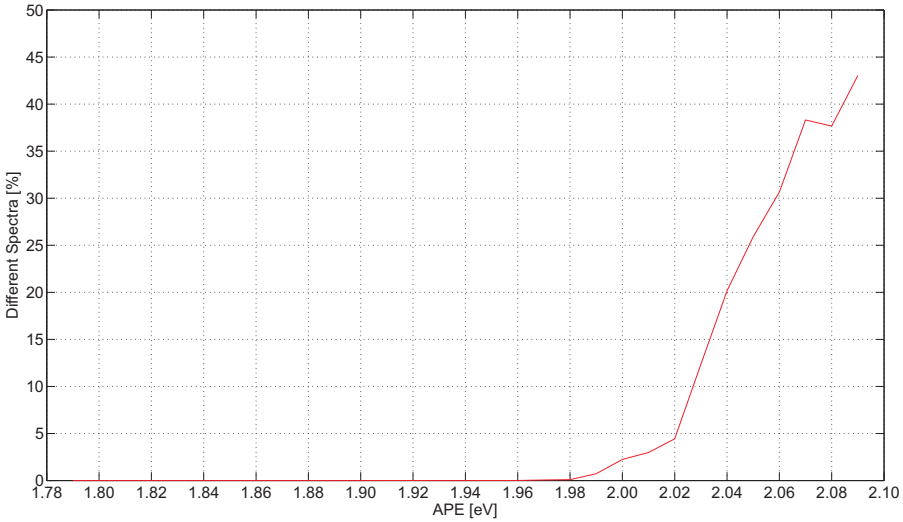

Fig. 10. Percentage of cumulative probability distribution functions of spectra which are the different for each APE value.

5. Results

5.1. Equality of solar spectral irradiance distributions with the same APE

Using Eq. (7) there are a total of 31 groups of APE. The statistic of Eq. (5) has been estimated for every pair of c.p.d.f. in each group. The level of confidence α is set at 0.05. If the difference $dist_{n,m}$ (Eq. (5)) is lower than the statistic c_α that corresponds to this significance level the two c.p.d.f.'s are the same for each comparison otherwise they are different. Fig. 10 shows the percentage of cumulative probability distribution functions that can be considered equals for each value of APE ($APE \pm 0.05$).

When analyzing the results in Fig. 10, it can be observed that almost all the spectral irradiances are the same that those of their class for APE values lower than 2.00 eV, that represent more than the 98% of spectra. Hence we can conclude that the spectra from 1.79 to 2.00 eV are the same to those with the same APE value.

5.2. Types of solar spectral irradiance distributions

For this analysis we have only used the spectra with APE ranges from 1.78 to 2.00 eV as the spectral irradiance distributions for greater values are different even for the same APE, see Fig. 10.

Using *k*-means data mining technique, all the spectra have been clustered using different number of cluster. After assigning the spectra c.p.d.f. to cluster and the distance of each sample to its cluster is performed to verify that the clustering is reliable. The confidence level is set at 0.05. Table 1 shows the results obtained when 2, 3, 4, 5 and 6 different clusters are used. Third column shows the percentage of elements in each cluster for each one of the total clusters used and fourth column shows the percentage of spectra that have the same spectra probability distribution in each cluster.

Regarding this table, we have selected five clusters as the most homogeneous clustering option such that a similarity of 99.98% arises for every cluster; that is, only less than 0.02 per cent of the spectra are not similar to the centroid of their cluster, see Table 1. Fig. 11 shows how the values of APE are distributed in each cluster.

It is remarkable that there is a strong relation between the APE value and the cluster to which the corresponding c.p.d.f. belongs to, even though some spectra with the same APE value are allocated in different clusters. This happens for samples whose APE value is at the limit of values belonging to a cluster. For instance, for APE values in the range [1.91 – 1.92), the percentage of samples in cluster 3 is 5.07 and in cluster 4 is 5.62. For these cases, the c.p.d.f. that corresponds to the APE value could be either three or four clusters. These measurements could be assigned to both clusters using rough *k*-means clustering, (Chen and Miao, 2011); it could be an improvement of the *k*-means used but this is beyond the scope of this paper where we are more interested in obtaining a general and easy method to classify each spectrum.

Using the proposed clustering, we have analyzed the relationship between solar spectral irradiance distribution and the performance of different modules and how the selected centroids can explain these performances.

Table 1
Spectra that meet the requirements of closeness to its centroid (percentage).

# Clusters	# Centroid	Spectra per cluster (%)	CDPF equal per (%) cluster
2	1	32.92	99.45
	2	67.08	97.60
3	1	25.23	99.66
	2	63.48	100.00
	3	11.29	98.67
4	1	26.36	100.00
	2	9.33	99.10
	3	52.70	100.00
	4	11.61	99.78
5	1	4.42	99.98
	2	48.40	100.00
	3	13.92	100.00
	4	23.52	100.00
	5	9.74	99.81
6	1	11.02	100.00
	2	40.70	100.00
	3	5.93	99.90
	4	3.74	100.00
	5	23.19	100.00
	6	15.43	100.00

5.3. Performance of thin-film photovoltaic modules

The types of c.p.d.f. found using *k*-means allows us to better explain the contour graphs previously presented, Fig. 1: If we analyze the obtained data for the temperature at STC (25°), the changes in

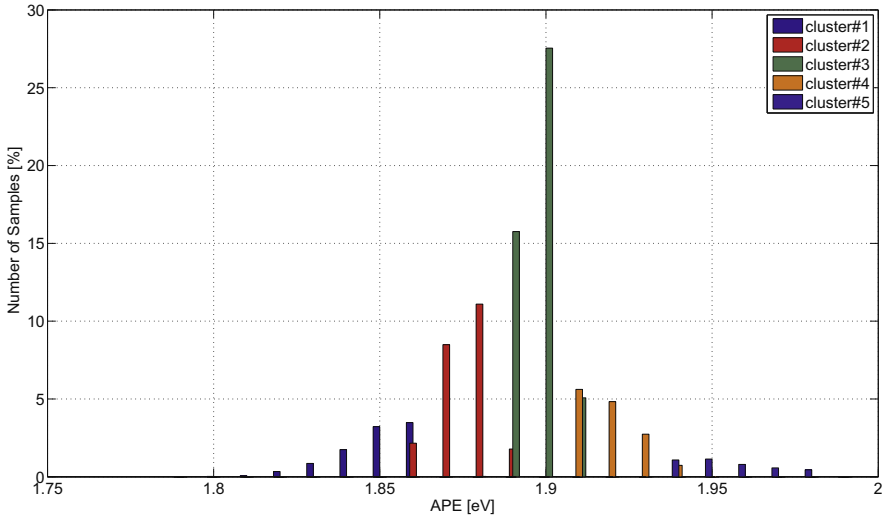

Fig. 11. Distribution of APE values in each cluster when 5 different clusters are used.

the value of PR are well explained if we consider the changes in the solar spectral irradiance distribution for the different APE values. That is, it is not enough to use only the values of APE to understand the changes in PR as the same PR value corresponds to different APE values. However, if we add the changes in the spectral distribution to the analysis, there is a clear correspondence between the cluster to which the APE belongs to and the PR value for this APE.

For a-Si modules, at 25° the PR ranges from 1.83 to 1.93 eV. The lower PR values (less than 85%) correspond to the APE values that belong to the first cluster; PR values that range from 85% to 90% correspond to the measurements included in the second cluster (APEs from 1.85 to 1.88 eV); PR values between 90% and 95% correspond to measurements that belong to the third cluster where APE values are in the 1.89–1.90 eV interval; PR values from 95% to 100% are estimated using APE values that range from 1.91 to 1.92 eV; finally, the greatest PR values (more that 100%) are obtained for the measurement in the fifth cluster whose APE values are greater than 1.91 eV.

Similar results can be observed in the other two technologies. For the same temperature, 25°, CdTe has PR under 90% for APE values lower than 1.82 eV; the PR varies from 90% to 95% for APE values that range from 1.82 eV to 1.88 eV (first and second cluster); the PR ranges from 95% to 100% for APE values between 1.88 to 1.90 eV (third cluster); PR values greater than 100% correspond to APE estimates for the measurements in the fourth and fifth clusters.

The PR for a-Si/μc-Si module also shows the same dependence on solar spectral irradiance distributions. The changes in the distribution from one cluster to another correspond to changes in PR values. PR values are under 90% for measurements whose APE value is below 1.88 (first and second cluster); there are PR values that range between 90% to 100% for APE values that range from 1.88 to 1.91 eV (third cluster); the highest values of APE (greater than 1.91 eV) correspond to PR greater than 100%.

We can conclude that there is a significant relation between the spectral distribution clusters proposed and the changes observed in the performance ratio of the modules. As a matter of fact, it is important when modeling the PR of a module to consider not only the solar irradiance received and temperature but also the type of spectral distribution that corresponds to the irradiance received. An important contribution of this paper is that it is possible to model all the solar irradiance spectral distributions measured in Malaga using only 5 different types of distributions. It would be interesting to establish whether this is the case for other different type of climates.

As can be noted from these figures, optimal conditions are achieved when modules reach PR close to 100%. As the peak power is given in STC of 25°C and AM 1.5 (1.88 eV), the PR will be the highest when the modules are working at around this values. A PR of 105% can be seen, as optimal conditions are surpassed and an APE of over 1.88 eV at a temperature of 25°C has occurred.

A very interesting result given by the three contour graphs intensifies the idea of clustering all APE in five groups. Although all three modules belong to different cell technologies and have different behavior for different outdoor conditions, all the changes in the PR can be explained using the selected clusters. If we focus on one fixed module temperature, for example the STC temperature, and we look at the different APE values at this temperature interval, we can see that APEs gather into different colors. These colors are related to centroids and APEs are clustered in every centroid.

6. Conclusions

We have analyzed the influence of the solar spectral irradiance distribution on the performance ratio of thin-film photovoltaic modules of different technologies. To carry out this analysis, we first characterize the cumulative probability distribution function of each spectrum using its APE value. After analyzing more than two hundred and fifty thousand measurements of solar spectral irradiation distributions, we can conclude that all these distribution can be represented using only 5 different distributions related to the APE values. We use statistical and data mining techniques to state this fact. Secondly, we prove that all the spectra that have the same APE value also have the same distribution shape. We here use the Kolmogorov–Smirnov two sample test that allows us to compare whether or not two cumulative probability distributions are the same. After perform this, we find that there are only five different types of distributions using the *k*-means data mining technique. Thanks to this technique it has been possible to reduce over two hundred and fifty thousand measured spectra to only five. Moreover, these five spectra are selected using the *k*-means algorithm; in fact, they are the centroids of each cluster. Almost all the spectra, more than 99.98% included in each cluster are the same to the centroid of the cluster using statistical criteria.

Using these five selected distributions, the performance ratio values estimated for photovoltaic modules built with different technologies are analyzed. A detailed analysis of the curves of performance ratio obtained when they are used as independent variables the temperature and the APE value for each outdoor

conditions allows us to conclude that the changes in the performance ratio of modules can be better explained if it is considered the spectrum associated to each APE value.

All these results have been obtained using data from Malaga. It would be of great interest to perform them for other locations with different climatic conditions.

Acknowledgments

This work has been supported by the projects P10-TIC-6441 and P11-RNM-07115 of the Junta de Andalucía, Spain.

References

ASTM G173-03. (2012). Standard tables for reference solar spectral irradiances: Direct normal and hemispherical on 37° tilted surface.

Berman, D., Faiman, D., & Farhi, B. (1999). Sinusoidal spectral correction for high precision outdoor module characterization. *Solar Energy Materials & Solar Cells, 58*, 253–264.

Cao, F., & Liang, J. (2011). A data labeling method for clustering categorical data. *Expert Systems with Applications, 38*, 2381–2385.

Chen, M., & Miao, D. (2011). Interval set clustering. *Expert Systems with Applications, 38*, 2923–2932.

Duda, R., Hart, P., & Stork, D. (2001). *Pattern classification*. John Wiley & Sons.

Fabero, F., & Chenlo, F. (1991). Variance in the solar spectrum with the position of the receiver surface during the day for PV applications. In *Proceedings of the 22nd IEEE photovoltaic specialists conference* (pp. 812–817). New York: IEEE Press.

Gottschalg, R., Infield, D. G., & Kearney, M. J. (2003). Experimental study of variations of the solar spectrum of relevance to thin film solar cells. *Solar Energy Materials & Solar Cells, 79*, 527–537.

Han, J., & Kamber, M. (2001). *Data mining concepts and techniques*. San Francisco: Morgan Kaufmann.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference and prediction*. Springer.

IEC 60904-3:1989, Edition 2.0 2008–04. Photovoltaic devices – Part 3: Measurement principles for terre strial photovoltaic (PV) solar devices with reference spectral irradiance data.

IEC 61215-2:1993, Second edition 2005–04. Crystalline silicon terrestrial photovoltaic (PV) modules – design qualification and type approval.

IEC 60904-7:1995, Edition 3.0 2008–11. Photovoltaic devices – Part 7: Computation of the spectral mismatch correction for measurements of photovoltaic devices.

Ishii, T., Otani, K., Takashima, T., & Xue, Y. (2011). Solar spectral influence on the performance of photovoltaic (PV) modules under fine weather and cloudy weather conditions. *Progress Photovoltaics: Research & Application*. http://dx.doi.org/10.1002/pip.1210.

Jain, A., Murty, M., & Flynn, P. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264–323.

Kymakis, E., Kalykakis, S., & Papazoglou, T. M. (2009). Performance analysis of a grid connected photovoltaic park on the island of crete. *Energy Conversion and Management, 50*, 433–438.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth symposium on math, statistics, and probability* (pp. 281–297). Berkeley, CA: University of California Press.

Martín, N., & Ruiz, J. M. (1999). A new method for the spectral characterisation of PV modules. *Progress Photovoltaics: Research & Application, 7*, 299–310.

Minemoto, T., Nagae, S., & Takakura, H. (2007). Impact of spectral irradiance distribution and temperature on the outdoor performance of amorphous Si photovoltaic modules. *Solar Energy Materials & Solar Cells, Japan, 91*, 919–923.

Minemoto, T., Nakada, Y., Takahashi, H., & Takakura, H. (2009). Uniqueness verification of solar spectrum index of average photon energy for evaluating outdoor performance of photovoltaic modules. *Solar Energy, 83*, 1294–1299.

Nann, S., & Emery, K. (1992). Spectral effects on PV-device rating. *Solar Energy Materials and Solar Cells, 27*, 189–216.

Nann, S., & Riordan, C. (1991). Solar spectral irradiance under clear and cloudy skies: Measurements and a semiempirical model. *Journal of Applied Meteorology, 30*, 447–462. http://dx.doi.org/10.1175/1520-0450(1991)030<0447:SSIUCA>2.0.CO;2.

Piliougine, M., Carretero Rubio, J. E., Mora Lopez, L., & Sidrach de Cardona Ortín, M. (2011). Experimental system for current–voltage curve measurement of photovoltaic modules under outdoor conditions. *Progress in photovoltaics*, 1–12.

Poissant, Y., Couture, L., Dignard-Bailey, L., Thevenard, D., Cusack, P., & Oberholzer, H. (2003). Simple test methods for evaluating the energy rating of PV modules under various environmental conditions. In *Proceedings of ISES 2003*, Gothenburg, Sweden.

Reda, I., & Andreas, A. (2004). Solar position algorithm for solar radiation applications. *Solar Energy*, 0038-092X, *76*(5), 577–589.

Rohatgi, V. K., & Saleh, A. K. M. E. (2001). *An introduction to probability and statistics* (second ed.). New York: Wiley-Interscience.

Williams, S. R., Betts, T. R., Helf, T., Gottschalg, R., Beyer, H. G. & Infield, D. G. (2003). Modelling long-term module performance based on realistic reporting conditions with consideration to spectral effects. In *Proceeding of the third world conference on photovoltaic energy conversion* (pp. 1908–1911). Osaka, Japan.

Zanesco, I., & Krenzinger, A. (1993). The effects of atmospheric parameters on the global solar irradiance and on the current of a silicon solar cell. *Progress Photovoltaics: Research & Application, 1*, 169–179.