

A data mining system for predicting solar global spectral irradiance. Performance assessment in the spectral response ranges of thin-film photovoltaic modules

J. del Campo-Ávila^a, M. Piliouginé^a, R. Morales-Bueno^a, L. Mora-López^{a,*}

^a*Departamento de Lenguajes y Ciencias de la Computación, ETSI Informática
Universidad de Málaga. Campus de Teatinos. 29071 Málaga. Spain*

Abstract

Knowing the spectral distribution of solar radiation is required to estimate the performance of photovoltaic modules, especially for thin-film modules. This is not a trivial problem due to the large number of environmental factors that affect this distribution as solar radiation passes through the atmosphere. The use of techniques of artificial intelligence and data mining can help in the development of models to address this problem. A system based on these techniques is proposed to predict the solar global spectral irradiance requiring only a few meteorological variables as inputs. The evaluation of the proposed system has been carried out for different wavelengths taking into account the spectral response of different technologies of thin-film photovoltaic modules. The errors in predicting solar global spectral irradiance for wavelengths that range between 350 and 900 nm and air mass lower than 2.1 are smaller than 7% on clear-sky days and than 16% for cloudy days, which is a significant improvement on other proposed models. Moreover, an open access implementation of the developed system is available at the URI: <http://fvred1.ctima.uma.es>. It could be useful for engineers and companies in the fields of the environment and renewable energies.

Keywords: solar spectral irradiance, data mining system, random forest, decision trees, open access software

*Corresponding author *Email address:* llanos@uma.es (L. Mora-López)

This is the accepted manuscript version submitted to Renewable Energy (17 September 2018)

Available online since 22 October 2018 at <https://doi.org/10.1016/j.renene.2018.10.083>.

Date of publication: April 2019

License: Attribution-NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

1. Introduction

Renewable energy has emerged as an increasingly competitive way to meet new power generation needs. Incidentally, the serious problems posed by climate change means that renewable energies are called on to play an increasingly important role in the current energy mix. According to the data included in the Report published by the International Energy Agency (Report IEA, 2017), in 2016 a total of 75 GW of photovoltaic energy (PV) were installed in the world, representing an increase of 50 % on the power installed in 2015.

This growth in the number of installations connected to the electricity grid poses an important challenge in terms of its correct integration in the electricity system: the prediction of its production. Therefore, forecasting the power that would be produced by photovoltaic plants is a matter of interest but is not a straightforward problem as this power depends on the availability of the solar resource, and it is difficult to predict. In addition to the main influential parameters to determine the performance of a PV module (irradiance and cell temperature), solar spectral distribution is another important factor, mainly when modules of solar thin-film technologies are used [1, 2, 3].

The different gasses in the Earth's atmosphere does not affect all types of photons in the same way and some wavelength bands experiment a significant reduction. Therefore, the solar spectrum presents a high variability with location and time. Two kinds of models to estimate the spectrum at Earth's surface can be found in the literature. On the one hand, there are radiative transfer methods, which are complex and rigorous [4, 5, 6]. They take into account measured vertical profiles of the layers of the atmosphere, which constitute a massive dataset. Consequently, they require high computational resources and large execution times.

On the other hand, the atmospheric transmittance methods are simpler models where each physical phenomenon that occurs in the atmosphere is modelled by a simple formula. These expressions are combined to synthesise the shape of the spectrum at specific locations and conditions. Bird presents a very simple model to estimate the solar spectrum requiring minimum computational resources, [7]. It is based on several previous works, mainly the papers by [8] and [9]. In the subsequent work by [10] the SPCTRAL2 model is described incorporating several improvements to estimate the diffuse spectral irradiance taken from [11]. Gueymard presents SMARTS2 [12], a simple

radiation model to estimate the spectrum for cloudless atmosphere for every plane orientation that outperforms previous models, especially when the zenith angle is high. However, according to [13], the success of these methods depends on the availability of certain atmosphere indexes that are hard to find for a specific location, which makes it difficult to apply for photovoltaic applications.

Data mining techniques can be incorporated to improve these models. In the research by [14], a complex physical model is employed repeatedly to simulate the spectral irradiance for 153 discrete wavelengths points from 280 to 700 nm for different combinations of atmospheric conditions. Then, a multilayer perceptron with 153 neurons in the output layer is trained with this dataset. Once the neural network has been trained, the solar spectrum for specific conditions can be obtained with minimum computation time. [15] have developed another neural network model to estimate the solar spectrum that is also valid for covered skies. For each discrete wavelength point from a set of 66 selected values (from 300 to 1100 nm), a different neural network is trained to obtain the spectral irradiance at that wavelength, using the spectral irradiance value for a cloudless sky (using the SPCTRAL2 model [10]), the air mass, and the global and direct clearness indices as inputs. [16] present a multilayer perceptron to obtain the spectral irradiance distribution using only the horizontal global irradiance, the air temperature, the air mass and the clearness index as inputs. In addition, a self-organised map was used in order to perform a selection from the most representative samples from the original dataset improving the generalisation power of the neural network (this selection technique was previously used in a work by [17]). In a paper by [18], a statistical analysis is performed on a dataset of experimental spectra measured over one year. They conclude that all these spectra can be classified according to their shapes into a few clusters, each one characterised by a representative spectrum (its centroid) and its APE (average photon energy), a value that can be calculated from the spectrum itself. All the spectra of the same type are very similar and only differ by a scaling factor. In a later paper, [19] study a way to obtain the solar spectrum using only a few meteorological parameters that are easily available at every weather station.

Our aim is to build a system that helps solar engineers to forecast the solar spectrum based on a reduced number of meteorological magnitudes (that can be easily measured at surface level using low cost instruments) and solar astronomical relationships (such as the sun elevation angle, which can

be accurately accessed from a particular location, date and time). We seek to solve this problem by combining the use of several techniques with the aim of achieving a more accurate prediction of solar spectra. In other contexts, like prediction of air and dew temperature, where meteorological variables are used as inputs, the use of combined data mining models improves results obtained too, [20].

The rest of the paper is organized as follows: Materials and methods are detailed in Section 2. The proposed methodology is described in Section 3. The description of the used dataset to train the models is provided in Section 4. A discussion of the results obtained when comparing the measured and predicted spectra using the different proposed models is presented in Section 5. The conclusions of this work are summarised in Section 6. The description of the implemented open access software that uses the best trained models is presented in Appendix A. It can be used to generate solar spectra providing only a few meteorological parameters.

2. Materials and methods

This section presents briefly the input parameters used to characterise and to predict solar global irradiance spectra. We then describe the basis of the data mining models used in the system (all of them induced by implementations available in the Weka framework [21]). Finally, we enumerate some metrics and methods used to estimate the performance of analyzed models.

2.1. Expression to calculate the atmospheric parameters

In addition to the meteorological parameters normally used in spectra characterisation and prediction, which are described in Section 4, the following atmospheric parameters were used as independent variables in the different models analysed:

- Air mass, AM , that is estimated using the following expression [22]:

$$AM = \frac{1}{\sin \alpha + 0.50572(\alpha + 6.07995)^{-1.6364}} \quad (1)$$

where α is the solar elevation (expressed in degrees). The coefficients in this expression were estimated using numerical data obtained from the ISO Standard Atmosphere model (ISA) that are valid for the measurements used in our study. The solar elevation is determined by [23]:

$$\alpha = \arcsin(\sin \delta \sin \phi + \cos \delta \cos \phi \cos \omega) \quad (2)$$

while δ is the Earth's declination, ϕ the latitude and ω the local hour angle.

- Clearness index, K_t , that is estimated using [23]:

$$K_t = \frac{G_t}{G_0} \quad (3)$$

where G_t is the measured horizontal global irradiance and the solar extraterrestrial irradiance, G_0 , is calculated as [23]:

$$G_0 = I_{sc} E_0 \sin \alpha \quad (\text{Wm}^{-2}) \quad (4)$$

while I_{sc} is the solar constant (1367 Wm^{-2}) and E_0 is the eccentricity correction factor. This factor is estimated as [24]:

$$E_0 = 1.000110 + 0.034221 \cos \Gamma + 0.001280 \sin \Gamma + 0.000719 \cos 2\Gamma + 0.000077 \sin 2\Gamma \quad (5)$$

where Γ is the day angle that is estimated using the day number of the year d_n ($1 \leq d_n \leq 365$):

$$\Gamma = \frac{2\pi(d_n - 1)}{365} \quad (6)$$

2.2. Classification models

The classification models analyzed to deal with non-numerical attributes are as follows: ZeroRules and Decision Stump (considered as the base line of the classifiers), Naïve Bayes, IB1 (a k-nearest neighbour model that only consider the closest example), MLP (multi-layer perceptron), J48 (implementation of C4.5, a decision tree), VFDT (an incremental decision tree algorithm particularly appropriate for large datasets) and Random Forest (an evolution of bagging that includes the ability to reduce overfitting most of the time). A detailed description can be found in [25]. They are succinctly described in the following subsections.

2.2.1. Naïve Bayes

This algorithm is one of the simplest variants using the Bayes Rule [26]:

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)} \quad (7)$$

It seeks to estimate the posterior probability $P(\text{class} | \text{observation})$ of each class given an observation and then selects the most likely class. Although the attributes must supposedly be conditionally independent, this algorithm performs very well even when that assumption does not occur. It can deal either with categorical or numerical attributes and it is robust in the presence of noise or missing attributes.

2.2.2. K -nearest Neighbour

The instance-based algorithms are another very simple approach, but that simplicity does not imply poor performance. They are based on determining which examples in the dataset are the most similar to the new observation. The output of the algorithm will take in consideration such information and calculate a distance function. One of the most commonly used methods is the k -nearest neighbour (where k refers to the number of neighbours to consider) [27]. One of the objections to these methods is the computational cost that depends on the dataset size as the observation must be compared against all the examples in the dataset.

2.2.3. Multilayer perceptron

A multilayer perceptron (MLP) is a feedforward artificial neural network that generates a set of outputs from a set of inputs. An MLP is characterized by several layers of input nodes connected as a directed graph between the input and output layers. It is trained using the Levenberge-Marquardt back-propagation algorithm, in which the artificial neural networks are organised in layers and send their signal forward, and the errors are propagated backwards. This learning function uses an adaptive learning rate. The behaviour of this type of network depends on multiple factors, such as the transfer function and the number of neurons in the hidden layer.

2.2.4. Decision trees

The algorithms that induce decision trees are widely used and multiple alternatives likewise exist. Some of their characteristics are particularly noteworthy, such as the capacity of being easily understandable (even by

non-expert persons) and the ability of recursively partitioning the problem in order to simplify its resolution (in a divide-and-conquer strategy). One of the most well-known algorithms is C4.5 [28], which allows the user to work with numerical data (besides categorical data), missing attributes, etc. The algorithm for building trees employs a top-down, greedy search through the space of possible branches. It uses entropy and information gain to generate a decision tree.

The entropy is used to calculate the homogeneity of a sample. It is necessary to calculate both the entropy before splitting and after splitting. The expression to calculate the entropies is shown in eq. 8.

$$H(X) = - \sum_{i=1}^n p(x_i) \log_b p(x_i) \quad (8)$$

where p is the frequency of attribute X with n outcomes, and b is the base of the logarithm, usually equal to 2.

The information gain, $I(X, Y)$, is based on the decrease in entropy after a dataset is split on an attribute X . It is calculated using eq.9.

$$I(X, Y) = H(X) - H(X|Y) \quad (9)$$

In addition to the classical batch learning procedure to induce the decision tree to consider the whole dataset (as is the case of C4.5), the decision tree can be induced in an incremental way, which allows better planning of the resources, particularly when datasets are large. One of the most used algorithms is VFDT [29], which uses Hoeffding's concentration bounds before selecting the best attribute to split a node.

2.3. Models for regression

The proposed methodology requires the use of regression models to deal with numerical attributes. The methods selected are Linear regression (considered as the base line model), M5P and REP Tree (regression trees) and Random Forest. In the following subsections the main ones are briefly described.

2.3.1. Linear regression

Linear regression is one of the most commonly used type of regression models. It works with experiences characterized by numeric attributes. Linear regression is a geometric method to approximate a cloud of points (the

experiences with n attributes) by means of a linear equation in a space of n dimensions (the number of attributes):

$$a_1 * x_1 + a_2 * x_2 + \dots + a_n * x_n = a_0 \tag{10}$$

where all the exponents are 1.

A linear equation represents a subspace of dimension $n - 1$ known as a hyperplane. The linear regression method obtains values for a_i , with $i = 0, \dots, n$ in the linear equation, that is, the better approximation for the cloud of points (from Machine Learning point of view, the linear equation is an explanation of the set of experiences).

2.3.2. *M5' (or M5P)*

M5' is an evolution of M5 algorithm defined by [30]. It was proposed by [31] and it is implemented as M5P in the data mining platform called Weka [21]. Let us briefly describe M5. It was defined to predict values rather than categories. The model tree construction is based on the divide-and-conquer method. The training set is split by all possible tests. The error is measured by considering standard deviation. The test that maximise the error reduction is chosen. The major innovations of M5 are:

1. Underestimating the error on unseen cases, by multiplying the error value by $(n + v)/(n - v)$, where n is the size of the training set and v the number of parameters. Therefore, the error increases if there are many parameters and a small number of cases.
2. Using a linear model in nodes where standard regression techniques are used, but only considering the attributes in the subtree of this node.
3. Simplifying the linear model by eliminating parameters to minimise its estimated error.
4. Pruning by considering the best between the linear model and the subtree model. If the linear model is chosen, the subtree is pruned and smoothed based on the number of cases in the branch, the predicted value and a smoothing constant are obtained.

M5P is an evolution of M5 that uses a standard method of transforming a classification problem into a problem of function approximation. By using conditional probability, the greatest approximated probability value is chosen as the predicted class. Moreover, the smoothing procedure is more complex and is based on the linear model involved in the leaves.

2.3.3. REPTree

REPTree is an algorithm from Weka. REPTree is a fast decision tree learner. For classification of numeric attributes, the algorithm first sorts all numeric fields in the data-set once, at the start of the run, and then uses the sorted lists to calculate the right splits in each tree node. The right split minimises the total variance. Related to non-numeric (discrete) attributes, it uses a regular decision tree with reduced-error pruning. Entropy is the measure considered. Missing values are dealt with by splitting the corresponding instances into pieces (as in C4.5).

2.4. Ensemble models for classification or regression

The previously exposed models (classification and regression) are needed to improve the two different subsystems of the global system that seeks to predict the solar global irradiance spectrum. The algorithm described so far to induce such models are used in a stand-alone way, that is, every algorithm induces a concrete model, but only one. Another active research area in supervised learning focuses on multiple classifier systems that were benefited from the idea of using a committee or ensemble of models to perform that tasks. Many approaches can be used to define a multiple-classifier system, but two of the most successful methods are bagging and random forest [32]. In general, the methods are highly precise, are robust to outliers and noise and do not overfit. However, the results are more difficult to interpret than when a single decision tree is considered.

Bagging [33] is a method that induces an ensemble of M classifiers by building M different datasets from the original one. Every “new” dataset is constructed by selecting the examples uniformly at random with replacement, and then a base classifier is induced by using such a dataset. Subsequently, all base classifiers predictions are combined with a voting method. The method

is shown in 1.

Input : The data set D

for $k \leftarrow 1$ **to** M **do**

- | A sample D_k is obtained from the dataset D by means of selecting the examples uniformly at random with replacement.
- | By considering D_k a base classifier C_k is induced.

end

Output: C_k classifiers, $k = 1, \dots, M$. The global prediction is obtained combining the M classifiers with a voting method.

Algorithm 1: Bagging algorithm.

Random forest [34] is another method, related to bagging, that induces a set of individual trees (no classifiers). The other main difference is the selection of attributes used to induce the trees, because not all attributes are considered: only a number N of attributes can be used in each node (in general, N is substantially less than the number of available attributes), It uses a total of M classifiers. With more details, 2 shows how performs this method.

Input : The data set D

for $k \leftarrow 1$ **to** M **do**

- | A sample D_k is obtained from the dataset D by means of bootstrap.
- | By considering D_k and a random selection of N attributes ($N < M$), a decision tree T_k is built.

end

Output: T_k trees, $k = 1, \dots, M$. The global prediction is obtained combining the M trees with a voting method.

Algorithm 2: Random forest algorithm.

2.5. Error metrics

The proposed models performances were assessed with quantitative tools, [35]. Specifically, the error metrics used are the relative mean absolute error ($rMAE$), the relative root mean square error ($rRMSE$), the mean bias error (MBE) and the correlation coefficient (ρ). The root mean square deviation ($RMSE$) as defined in [36] has also been used:

$$RMSD = \left[\frac{1}{n} \sum_{i=1}^n \left(\frac{\hat{y}_i - \bar{y}}{\bar{y}} \right)^2 \right]^{1/2} \cdot 100 \quad (\%) \quad (11)$$

Moreover, the Kolmogorov-Smirnov two samples test has also been used to compare the probability distribution function of different spectra. This experimental statistic is estimated using the expression:

$$dist_{m,n} = \max_{s \in \mathbb{R}} |cp\hat{d}f_1(x_s) - cp\hat{d}f_2(x_s)| \quad (12)$$

where m and n are the size of sample 1 and sample 2, respectively, and $cp\hat{d}f_1$ and $cp\hat{d}f_2$ are the estimated cumulative probability distribution functions of such samples.

This value have to be less than the theoretical value c_α (that depends of the significance level, α):

$$\left(\frac{n \cdot m}{n + m} \right)^{1/2} \cdot dist_{m,n} < c_\alpha \quad (13)$$

3. Proposed methodology

This paper proposes relevant improvements to the system presented by [19], which uses both atmospheric and meteorological inputs to estimate the real spectrum. We can identify two different subsystems in that system: the one that classifies an observation, and the one that estimates the normalisation factor for such an observation. We must remark the existence of a third aspect in the system, the one responsible for the calibration of the clustering itself (with three groups or clusters), but we will not modify it given its good performance.

Then, the system for the prediction of the spectrum is divided into two different stages according to the scheme shown in Figure 1.

First, the shape of the spectral distribution is determined by selecting one of three possible predetermined spectra. Actually, these spectra are the centroids of the three clusters (or groups) in which the total set of spectra have been previously classified [19].

The cluster selection procedure is performed taking into account the meteorological input variables. However, each cluster has thousands of spectra with similar shape but different heights. In order to determine the predicted

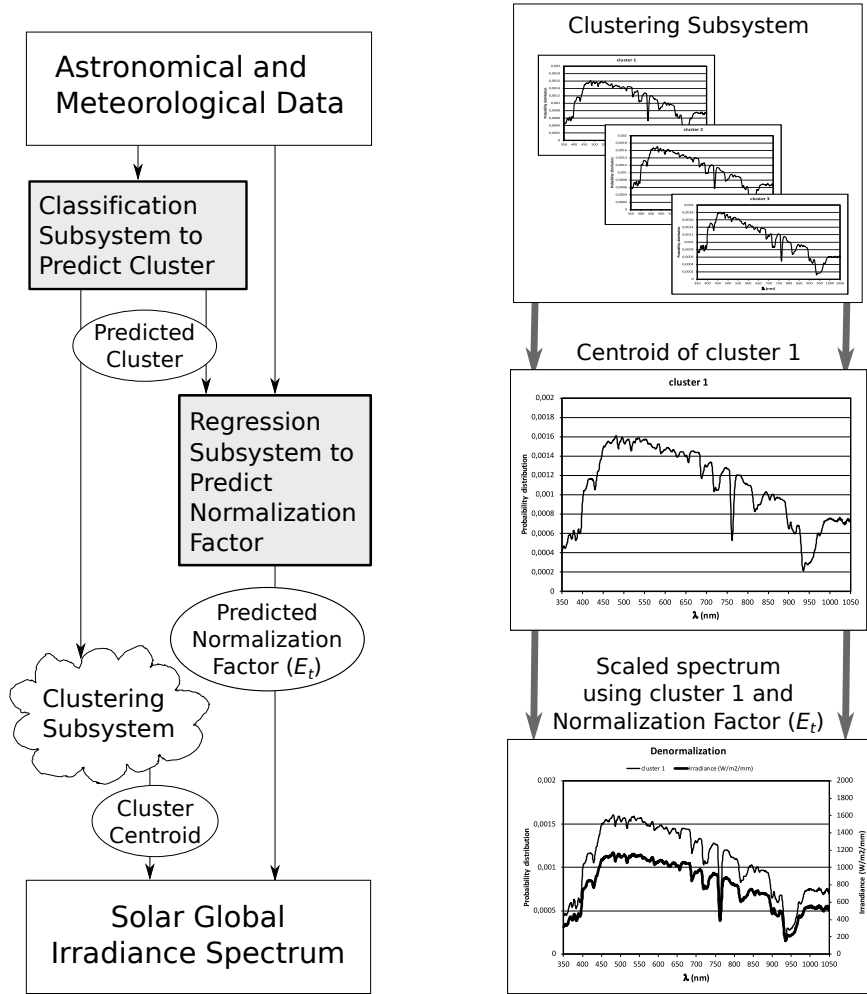


Figure 1: Prediction of a spectrum

spectrum with its actual height, the selected spectrum must be scaled by a normalization factor E_t . Therefore, the second task consists in computing this factor also using the meteorological input values.

Although the proposal to make the prediction in two stages is valid, we have detected that the proposed models for each stage can be improved. Several data mining models were analyzed in order to select the best model for each stage. For the first subsystem, classification models were checked while regression models were studied in the second subsystem. All the models were fitted using cross-validation.

4. Input data

Data were recorded at the Photovoltaic Systems Laboratory of the University of Málaga (latitude: 36.7 N, longitude: 4.5 W, altitude: 50 m). The meteorological data recorded are: air temperature, relative humidity, wind speed, horizontal global irradiance and the solar spectrum. These measurements were taken by a software running in a control computer which synchronizes several instruments. On the one hand, the meteorological conditions are acquired at regular intervals of time (every 5 minutes): the horizontal global irradiance (captured by a CMP21 pyranometer from Kipp and Zonen), the wind speed (sensed by an anemometer Young 3002L), the air temperature and the relative humidity (both of them readed from the combined probe Young 41382LC). All these sensors are connected to a programmable automation controller (National Instruments Compact FieldPoint cFP2120), whose registers are accessible at real-time through the protocol OPC-DA ([37]) over Ethernet. On the other hand, the spectrum of the solar radiation is captured using a grating spectroradiometer (EKO MS-710) connected to the control computer using a RS-485 bus. A proprietary protocol has been implemented to retrieve the solar spectrum at the same time the meteorological measurements are taken. The measurement equipment is calibrated every two years by an accredited laboratory (CIEMAT, Madrid).

The spectrum wavelengths range from 350 to 1050 nm which means that a total of 920 values were used for each spectrum. All the selected spectra were recorded under a solar elevation angle greater than 15° [38]. Measurements were collected in two different time periods. Data used to fit the models were measured from November 2010 to May 2012. A total of 265054 spectra was recorded. Data used to test the model were measured from March 2016 to May 2016. A total of 20292 measurements were used from this period. The values of Aerosol optical depth at 500 nm, Angstrom exponent and precipitable water values recorded for the period where measurements were recorded are shown in Table 1. These values have been obtained from the AERONET website (<https://aeronet.gsfc.nasa.gov>).

In addition to the meteorological data the following atmospheric parameters were used as independent variables: air mass (estimated as in [22]) and clearness index (estimated as in [23]). The mean values of atmospheric and meteorological data for each cluster of the dataset are shown in Table 2.

| Year | τ_{a500} | $\alpha_{440-870}$ | PW |
|------------------|---------------|--------------------|------|
| 2010 | 0.149 | 0.92 | 1.91 |
| 2011 | 0.155 | 0.99 | 2.01 |
| 2012 | 0.155 | 0.93 | 1.75 |
| 2009-2016 (mean) | 0.142 | 0.99 | 1.84 |

Table 1: Aerosol optical depth at 500 nm (τ_{a500}), Angstrom exponent ($\alpha_{440-870}$) and precipitable water (PW). Source: <https://aeronet.gsfc.nasa.gov>.

| Parameter | Fitting set | | |
|-------------------------------------|-------------|-----------|-----------|
| | Cluster 1 | Cluster 2 | Cluster 3 |
| Global solar irradiance (W/m^2) | 426 | 593 | 218 |
| Wind speed (m/s) | 2.2 | 2.1 | 2.1 |
| Relative humidity (%) | 52 | 51 | 67 |
| Outdoor temperature ($^{\circ}C$) | 18.2 | 24.8 | 19.4 |
| Air mass | 2.33 | 1.63 | 1.90 |
| Clearness index | 0.66 | 0.62 | 0.26 |

Table 2: Mean values of meteorological parameters for the samples in each cluster of the data used.

5. Results and discussion

5.1. Estimation of cluster

Obtaining the type of distribution of the spectrum is the first task to predict a new solar spectrum using meteorological data. Different classification algorithms were tested with different configurations. Weka was the framework used to fit the data mining models [21]. The most relevant results that support our final decision are summarized in Table 3. Taking into account these results, Random Forest is the method selected to be installed in the ‘‘Classification subsystem’’. The main reason is that it achieves the maximum accuracy, while maintaining a short time to predict an observation. These two characteristics are fundamental, because the system is going to be accessible via a web page (see Appendix A) and users will need to receive the best prediction as fast as possible. Other options such as IB1 offers good accuracy, but the prediction time is high (as the observation has to be compared against the entire dataset, approximately 265000 examples).

| Algorithm | % correct | Model size (KB) | Prediction time (ms) |
|----------------|-----------|-----------------|----------------------|
| ZeroR | 63.75 | 1 | 0.001 |
| Decision Stump | 63.78 | 2 | 0.001 |
| Naive Bayes | 76.64 | 3 | 0.006 |
| IB1 | 92.01 | 18371 | 45.360 |
| MLP | 82.32 | 12 | 0.003 |
| J48 | 90.58 | 2934 | 0.003 |
| VFDT | 83.01 | 165 | 0.004 |
| Random Forest | 92.95 | 244860 | 0.165 |

Table 3: Percentage of spectra equals to the centroid of the cluster

5.2. Estimation of normalization factor (E_t)

Each one of the four analysed regression methods was used to induce models for each cluster. The number of observations in each cluster was 66251 for Cluster 1, 168984 for Cluster 2 and 29819 for Cluster 3.

The correlation coefficient (ρ) and relative mean absolute error (rMAE) obtained (using cross-validation with 10 folds) are shown in Figure 2.

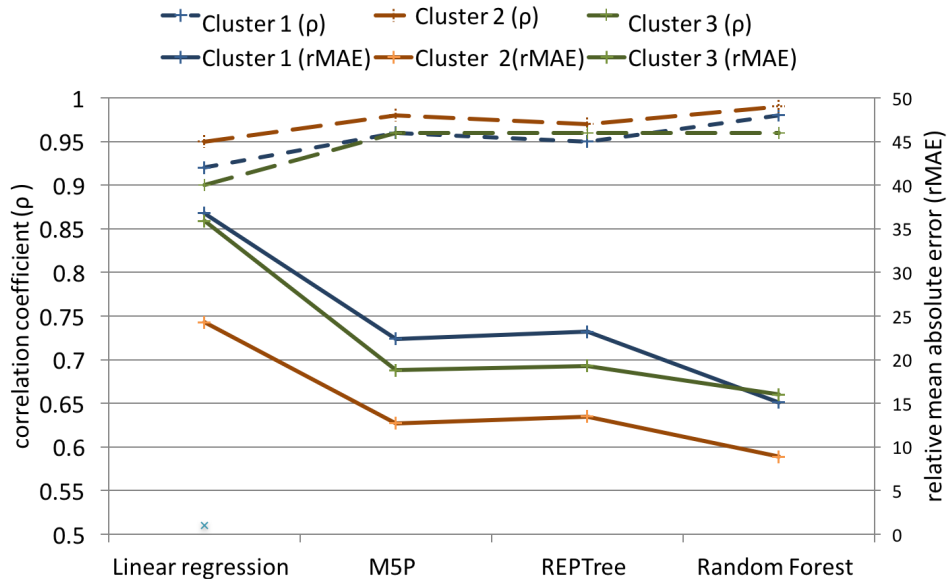


Figure 2: Error metrics in each cluster for the analyzed models to estimate the normalization factor

Random Forest is the model with lowest errors and greatest correlation coefficient for both experiments and for the three clusters. The correlation coefficients obtained for this new proposed model (0.98, 0.99 and 0.96 for Clusters 1, 2 and 3 respectively) improve significantly those obtained using linear regression proposed in [19] where the correlation coefficient obtained was 0.94. The following model with lower errors is M5P, which also improves the previously reported correlation coefficient. Both models will be used to estimate the total error of the prediction process of solar spectra using the above-mentioned meteorological data for samples not previously used.

5.3. Predictions and results for new observations

The proposed methodology has been validated with spectra that were not used to build the models. The following two tests have been conducted:

- One test after the first stage: the percentage of predicted solar spectral cumulative distribution that are statistically equal to the recorded ones.
- One tests after the second stage: the rMAE estimated for the prediction of the normalization factor

The number of records in the testing set is 20292. These new spectra have been recorded besides the meteorological data observed at the moment of recording. Thus, we can calculate the predicted spectra by using our proposed system and meteorological data, and it can be then compared with the measured spectra.

The cluster corresponding to each recorded spectrum has been obtained directly from the centroids proposed in [19] and the normalized spectrum using the Euclidean distance function. The number of recorded spectra of each cluster is 4276, 12824 and 3192 respectively for Clusters 1, 2 and 3.

Once all the information of recorded spectra (cluster and normalization factor) has been estimated, the following task is to predict new spectra from meteorological data. The first stage to predict a spectrum is to determine the cluster. The meteorological data are used as entry to the Random Forest model (Section 5.1) and the output is the cluster predicted. After this stage, a collection of normalized spectra is obtained. In order to check this first stage, the recorded (normalized) spectra are compared with the centroid of the cluster assigned to each one using the Kolmogorov-Smirnov two sample test. For a significance level equal to 0.05, the percentages of recorded spectra that are statistically equal to the centroid of the assigned cluster are 98.8,

98.5 and 96.1 for clusters 1, 2 and 3 respectively; and for a significance level equal to 0.01 these percentages are 99.4, 99.1 and 97.6 respectively.

Once the clusters have been estimated, the two models analyzed to estimate the normalization factor were Random Forest and M5P. The estimated root mean relative error for each model and cluster is shown in Table 4.

| Model | Cluster1 | Cluster2 | Cluster3 |
|---------------|----------|----------|----------|
| M5P | 29.2 | 6.6 | 13.3 |
| Random Forest | 33.1 | 9.3 | 14.1 |

Table 4: Relative root mean square error in test set for estimating the normalization factor E_t

As can be observed the smallest errors are obtained when the M5P model is used. However, the model that presented the smallest errors for data used in the adjustment was the Random Forest model. This can be explained by the overfitting that sometimes occurs when using Random Forest (especially when the number of attributes is small as happens in our dataset). Therefore, the M5P model was the model selected for estimating the normalization factor in the second subsystem of the developed tool (see 5.2) .

5.4. Comparison with other models

In order to compare the accuracy of the proposed model with some previous ones, the root mean square deviation (RMSD) defined in Section 2.5 has been used as the error measurement for comparing the measured spectral values and those obtained by the model. The values of wavelength used ranges from 350 to 1000 nm. The RMSD values obtained are 10.1 and 8.4 % for Random Forest and M5P respectively. These values improve the obtained in [36] that use SMARTS with data from Atenas (urban area) for predicting the spectral direct beam irradiance and the RMSD values range from 14 to 24 %. For that location, the aerosol optical depth at 500 nm is 0.174, the Angstrom exponent is 1.30 and the precipitable water is 1.68 for Atenas; these mean values have been obtained from the AERONET website (<https://aeronet.gsfc.nasa.gov>), from 2008 to 2017. Although Atenas (in Greece) and Málaga (in Spain) are obviously different cities, results are comparable because the atmospheric and meteorological conditions are, in effect, similar.

Finally, the mean differences between estimated and measured values (relatives to the measured values) and the mean bias error have been estimated

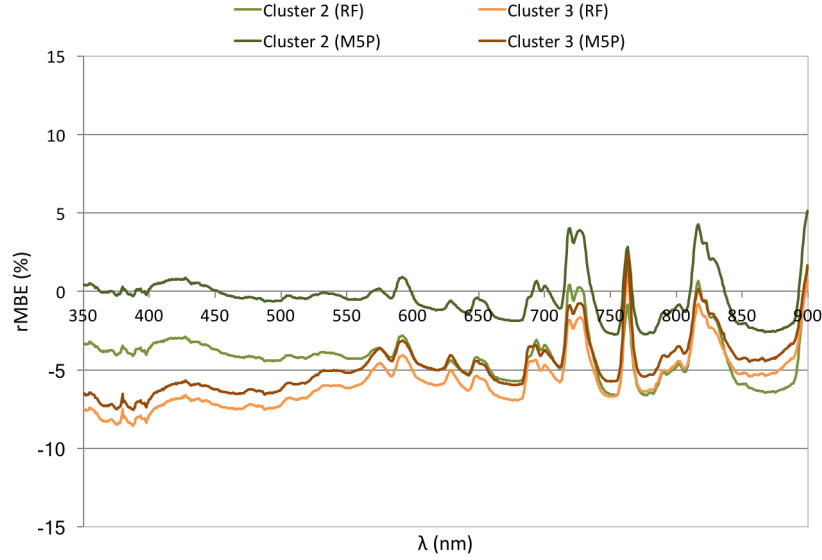


Figure 3: Mean bias error (%) of energy for each wavelength for Clusters 2 and 3 using Random Forest and M5P models.

for each wavelength, according to the proposal of [39]. Figure 4 shows the differences between estimated and measured values for Cluster 2 and Cluster 3 once the non-validity of the model for Cluster 1 has been shown. Figure 3 shows the values of mean bias error obtained (%) for Cluster 2 and Cluster 3.

In both cases, the observed differences are lower for M5P. For this model, these relative differences range between 5 and 10 % although for most of the wavelength are always lower than 6 %. These results improve the previously reported in [39] that range between -35 to 20 % whether considered wavelength vary between 300 and 1100 nm while for wavelength between 350 and 900 these differences vary between -12 and 5 %. The mean bias error ranges between -3 and +4 % for M5P model in cluster 2 and between -7 and 3 % in cluster 3; in this case, there is a tendency to underestimate the energy for most of the wavelengths.

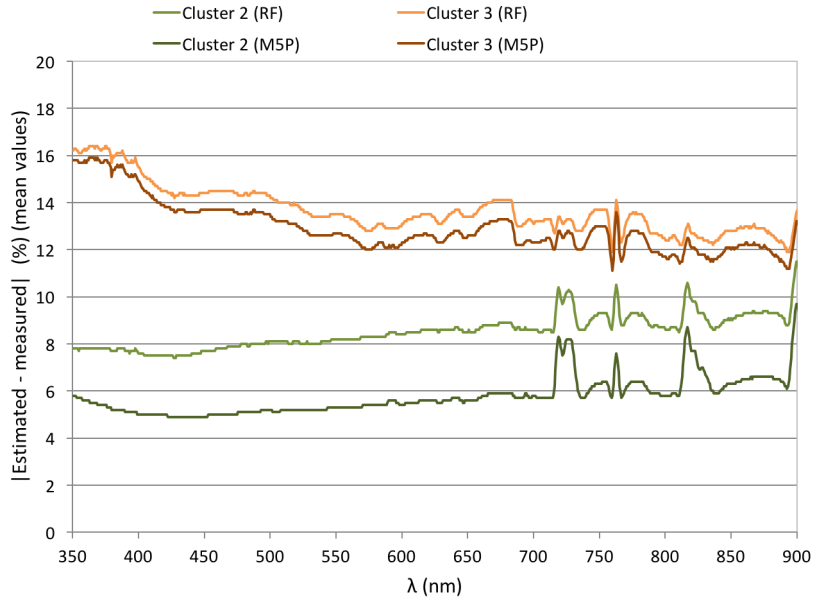


Figure 4: Relative differences between estimated and recorded values of energy for each wavelength for Clusters 2 and 3 using Random Forest and M5P models. $|A|$ means absolute value of A .

5.5. Evaluation of the proposed system in the spectral response ranges of thin-film photovoltaic modules of different technologies

As is well known, the performance of photovoltaic modules depends on the solar spectral distribution of the received solar radiation, especially for thin-film photovoltaic modules. The errors of the proposed system have been evaluated for several wavelength ranges taking into account the reported relative spectral response of several PV thin-film modules, [40]. [2], [41]. For instance, using the data published in [42], [43] and [44] the spectral response for different technologies is shown in Figure 5.

Some thin-film technologies, such as a-Si and CdTe, have spectral responses very narrow, as can be observed in Figure 5. Consequently, small alterations in the spectral distribution of sunlight can significantly affect the power output of modules of these technologies. In order to evaluate the possibility of using the proposed system to obtain this distribution and using it in the estimation of the performance of PV thin-film modules, the rMAE for the prediction of irradiance for 14 wavelength intervals has been estimated. The intervals of λ are as follows:

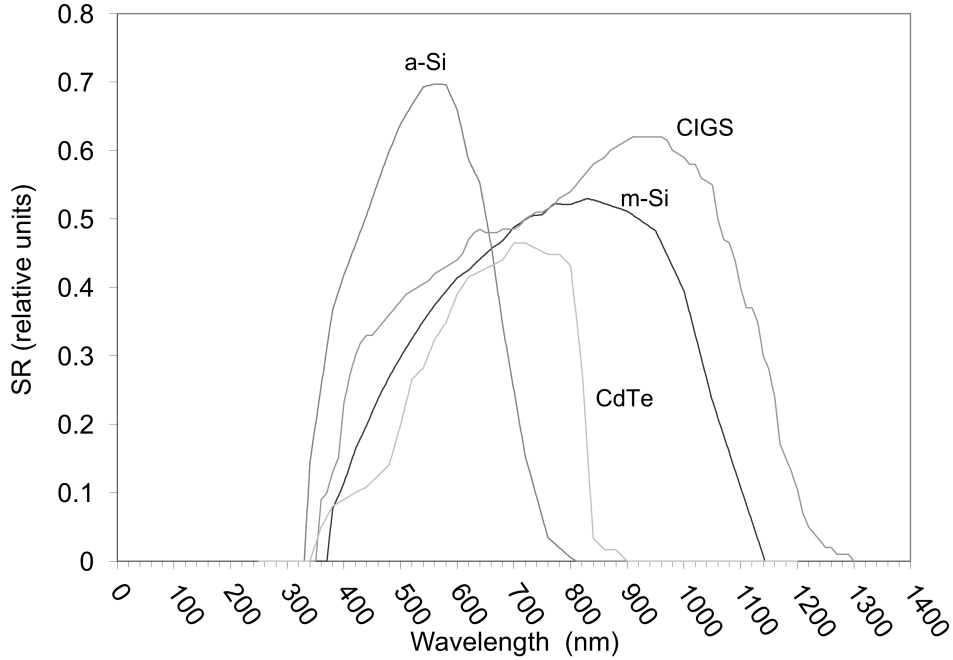


Figure 5: Relative Spectral response for different technologies (source [42] and [43]).

$$[350 + (i - 1) \cdot 50, 350 + i \cdot 50) \quad \text{for } i = 1, \dots, 14 \quad (14)$$

The results obtained are shown in Figure 6.

As can be observed for all the intervals and clusters the smallest errors are also obtained by the M5P model. Cluster 2, which corresponds to AM values of around of 1.6 and clearness index of 0.6, is where the errors are smaller; it is always less than 7 % for wavelengths less than 900 nm. This means that the model is able to predict the distribution of global solar spectral irradiance for clear days and central hours of the day with errors that improves significantly results obtained in the work of [19]. In Cluster 3 the errors for wavelength less than 900 nm are always less than 16 % which also improved the results obtained with the model proposed in [19]. This cluster corresponds to AM values of around 1.9 and the clearness index of around 0.25. This means that the proposed model is capable of making good predictions of the solar global radiation spectrum even for cloudy sky. The errors are significant only in Cluster 1. This cluster corresponds to AM of around 2.3 that means the proposed model should not be used for low solar heights (first and last hours

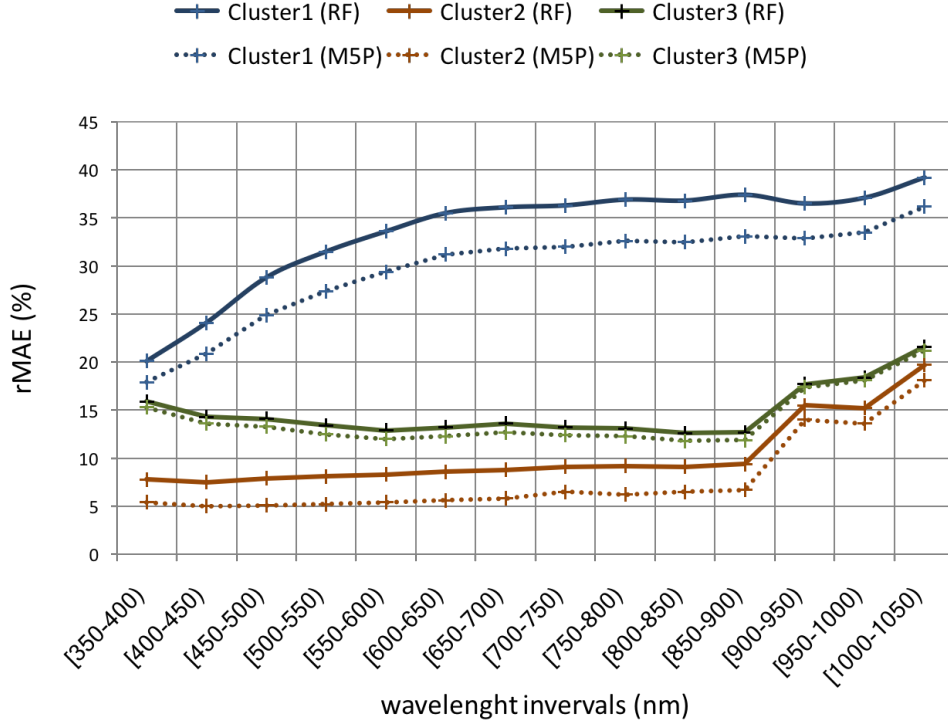


Figure 6: rMAE in each cluster for the two fitted models to obtain the normalization factor E_t . The cluster has been obtained in the first stage (see subsection 5.1) using Random Forest model.

in the day).

6. Conclusions

A methodology based on the use of certain classification and regression techniques to predict the solar global spectral irradiance distribution is proposed. Only several meteorological parameters are necessary to obtain the corresponding solar spectral distribution. The solar spectrum is obtained in two steps: in the first one, the type of spectrum that corresponds to the meteorological parameter is determined using a Random Forest model (classification technique); in the second one, the normalisation factor is estimated from these meteorological parameters using a M5P model (regression technique). This proposed methodology can be applied to any location. The current estimated models work with meteorological and atmospheric condi-

tions similar to data used in the training phase. Moreover, our system can be adapted to regions with very different conditions by refitting the models with new training sets.

The models have been evaluated for different wavelength ranges taking into account the spectral response of PV modules of different technologies. The errors in the prediction of solar global spectral irradiance for wavelengths that range between 350 and 900 nm and air mass lower than 2.1 are smaller than 7 % on clear-sky days and than 16 % for cloudy days. These air mass values correspond to the central hours of the day (for the latitudes used), when the received irradiance is greater.

The proposed methodology developed using data mining models improves previously reported results except for high values of air mass.

An open access software implementing the proposed models has been developed and is available at the URI: <http://fvred1.ctima.uma.es>. This software can be used in systems in which it is necessary to predict solar spectrum such as in the fields of energy, environment or agriculture.

As future research, it would be desirable to extend the proposed methodology so that it can be also used when only some of the meteorological data are available. In addition, the developed tool could also retrieve meteorological parameter information from external sources if it were available, without having to request that information from the user.

Acknowledgements

We acknowledge the *Universidad de Málaga* (Research Plan) and *Junta de Andalucía* (grant no. P11-RNM-7115) for financial support. We thank Vassilis Amiridis for his efforts in establishing and maintaining the ATHENS-NOA site; and CGL2016-81092-R and ACTRIS-2 (grant agreement N. 654109, European Union's Horizon 2020 research and innovation 810 programme) projects for establishing and maintaining the AERONET-Malaga site.

References

- [1] R. Gottschalg, D.G. Infield, and M.J. Kearney. Experimental study of variations of the solar spectrum of relevance to thin film solar cells. *Solar Energy Materials and Solar Cells*, 79(4):527 – 537, 2003.
- [2] Tetsuyuki Ishii, Kenji Otani, Takumi Takashima, and Yanqun Xue. Solar spectral influence on the performance of photovoltaic (pv) modules

- under fine weather and cloudy weather conditions. *Progress in Photovoltaics: Research and Applications*, 21(4):481–489, 2013.
- [3] Michel Piliouguine, David Elizondo, Llanos Mora-López, and Mariano Sidrach-de-Cardona. Photovoltaic module simulation by neural networks using solar spectral distribution. *Progress in Photovoltaics: Research and Applications*, 21(5):1222–1235, 2013.
- [4] F.X. Kneizys, G.P. Anderson, E.P. Shettle W.O. Gallery, L.W. Abreu, J.E.A. Selby, J.H. Chetwynd, and S.A. Clough. Users guide to lowtran 7. *Air Force Geophysics Laboratory*, 1010(AFGL-TR-88-04 77), 1988.
- [5] A. Berk, G.P. Anderson, P.K. Acharya, and E.P. Shettle. Modtran 5.2.0.0 user’s manual. *Spectral Sciences, INC. & Air Force Research Laboratory*, 1010(AFGL-TR-88-04 77), 1988.
- [6] Shepard A. Clough, Michael J. Iacono, and Jean-Luc Moncet. Line-by-line calculations of atmospheric fluxes and cooling rates: Application to water vapor. *Journal of Geophysical Research: Atmospheres*, 97(D14):15761–15785, 1992.
- [7] Richard E. Bird. A simple, solar spectral model for direct-normal and diffuse horizontal irradiance. *Solar Energy*, 32(4):461 – 471, 1984.
- [8] Bo Leckner. The spectral distribution of solar radiation at the earth’s surface—elements of a model. *Solar Energy*, 20(2):143 – 150, 1978.
- [9] D.T. Brine and M. Iqbal. Diffuse and global solar spectral irradiance under cloudless skies. *Solar Energy*, 30(5):447 – 453, 1983.
- [10] Richard E. Bird and Carol Riordan. Simple solar spectral model for direct and diffuse irradiance on horizontal and tilted planes at the earth’s surface for cloudless atmospheres. *Journal of Climate and Applied Meteorology*, 25(1):87–97, 1986.
- [11] C. G. Justus and M. V. Paris. A model for solar spectral irradiance and radiance at the bottom and top of a cloudless atmosphere. *Journal of Climate and Applied Meteorology*, 24(3):193–205, 1985.
- [12] Christian A. Gueymard. Parameterized transmittance model for direct beam and circumsolar spectral irradiance. *Solar Energy*, 71(5):325 – 346, 2001.

- [13] Marius Paulescu, Eugenia Paulescu, Paul Gravila, and Viorel Badescu. *Weather Modeling and Forecasting of PV Systems Operation*. Springer, 2013.
- [14] H. Schwander, A. Kaifel, A. Ruggaber, and P. Koepke. Spectral radiative-transfer modeling with minimized computation time by use of a neural-network technique. *Applied optics*, 40:331–5, Jan 2001.
- [15] A. Peled and J. Appelbaum. A solar spectrum model based on artificial neural-networks. In *29th European Photovoltaic Solar Energy Conference and Exhibition*, pages 2327 – 2334, 2014.
- [16] M. Torres-Ramírez, D. Elizondo, B. García-Domingo, G. Nofuentes, and D.L. Talavera. Modelling the spectral irradiance distribution in sunny inland locations using an ann-based methodology. *Energy*, 86:323 – 334, 2015.
- [17] M. Piliouline, J. Carretero, L. Mora-López, and M. Sidrach-de Cardona. Experimental system for current–voltage curve measurement of photovoltaic modules under outdoor conditions. *Progress in Photovoltaics: Research and Applications*, 19(5):591–602, 2011.
- [18] Rafael Moreno-Sáez, Mariano Sidrach-de-Cardona, and Llanos Mora-López. Data mining and statistical techniques for characterizing the performance of thin-film photovoltaic modules. *Expert Systems with Applications*, 40(17):7141 – 7150, 2013.
- [19] Rafael Moreno-Sáez and Llanos Mora-López. Modelling the distribution of solar spectral irradiance using data mining techniques. *Environ. Model. Softw.*, 53(C):163–172, March 2014.
- [20] Karthik Nadig, Walter Potter, Gerrit Hoogenboom, and Ronald McClendon. Comparison of individual and combined ann models for prediction of air and dew point temperature. *Applied Intelligence*, 39(2):354–366, 2013.
- [21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

- [22] F. Kasten and A. Young. Revised optical air mass tables and approximation formula. *Applied Optics*, 28(22):4735 – 4738, 1989.
- [23] Muhammad Iqbal. *An Introduction to Solar Radiation*. Academic Press, 1983.
- [24] J.W. Spencer. Fourier series representation of the position of the sun. *Search*, 2, 1971.
- [25] Claude Sammut and Geoffrey I. Webb. *Encyclopedia of Machine Learning*. Springer Publishing Company, Incorporated, 1st edition, 2011.
- [26] P. Domingos and M. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 291:103–130, 1997.
- [27] T. Cover and P. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13:21– 27, 1967.
- [28] J. R. Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann, 1993.
- [29] P. Domingos and G. Hulten. Mining high-speed data streams. In *Proceedings of the ACM 6th International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD-2000)*, pages 71–80, 2000.
- [30] Ross J. Quinlan. Learning with continuous classes. In *5th Australian Joint Conference on Artificial Intelligence*, pages 343–348, Singapore, 1992. World Scientific.
- [31] Y. Wang and I. H. Witten. Induction of model trees for predicting continuous classes. In *9th European Conference on Machine Learning (poster papers)*, pages 128 – 137. Springer, 1997.
- [32] Abraham J. Wyner, Matthew Olson, Justin Bleich, and David Mease. Explaining the success of adaboost and random forests as interpolating classifiers. *Journal of Machine Learning Research*, 18(48):1–33, 2017.
- [33] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.
- [34] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

- [35] N.D. Bennett, B.F. Croke, G. Guariso, J.H. Guillaume, S.H. Hamilton, A.J. Jakeman, S. Marsili-Libelli, L.T. Newham, J.P. Norton, C. Perrin, S.A. Pierce, B. Robson, R. Seppelt, A.A. Voinov, B.D. Fath, and V. Andreassian. Characterising performance of environmental models. *Environ. Modell. Softw.*, 40:1–20, 2013.
- [36] D.G. Kaskaoutis and H.D. Kambezidis. The role of aerosol models of the smarts code in predicting the spectral beam irradiance in an urban area. *Renewable Energy*, 33, 2008.
- [37] Opc data access custom interface specification, version 3.00. Technical report, OPC Foundation, 2003.
- [38] S. Nan and C. Riordan. Solar spectral irradiance under clear and cloudy skies: measurements and a semiempirical model. *Journal of Applied Meteorology*, 30:447 – 462, 1991.
- [39] Christian A. Gueymard. Smarts2, a simple model of the atmospheric radiative transfer of sunshine: Algorithms and performance assessment, 1995.
- [40] Paul Grunow, Alexander Preiss, Simon Koch, and Stefan Krauter. Yield and spectral effects of a-si modules. In *Proceedings of the 24th European Photovoltaic Solar Energy Conference*, pages 2846–2849, Hamburg, Germany, 2009.
- [41] Ira Devi Sara, Thomas R. Betts, and Ralph Gottschalg. Determining spectral response of a photovoltaic device using polychromatic filters. *IET Renewable Power Generation*, 8(5):467–473, 2014.
- [42] R. L. Mueller. The calculated influence of atmospheric conditions on solar cell isc under direct and global solar irradiances. In *Proceedings of the 19th IEEE PVSC*, pages 166–170, New Orleans, USA, 1987.
- [43] H. Field. Solar cell spectral response measurement errors related to spectral band width and chopped light waveform. In *Proceedings of the 26th IEEE PVSC*, pages 471–474, Anaheim, 1997.
- [44] J.J. Pérez-López, F. Fabero, and F. Chenlo. Experimental solar spectral irradiance until 2500 nm: Results and influence on the pv conversion of

different materials. *Progress in Photovoltaics: Research and Applications*, 15:303–315, 2007.

Appendix A. Open access software

The proposed model has been implemented in an open access software to be accessible to everyone with an internet connection. As the model has been fitted using Weka tools, it is possible to export a library implemented in Java that implements that model. The only encapsulation requirements of that library are the model input values in an object belonging to a Java class provided by the library itself. A code written in PHP gathers the input data from a web form filled by the user. The required object is created and supplied with that information by a wrapper also coded in PHP. The PHP-Java bridge has been used in order to allow the management of Java objects from PHP.

The developed software has been published on an open-access website (see Figure A.7).

REQUIRED INPUT VALUES:

Local time Universal time Solar time

Hour: Minute:

Day: Month: Year:

No daylight saving hour Apply daylight saving hour in summer (only EU)

Official time zone for this local time

Latitude (positive north of equator, negative south of equator) [degrees]

Longitude (positive east of Greenwich, negative west of Greenwich) [degrees]

Irradiance on horizontal surface [W/m²]

Wind speed [m/s]

Air temperature [°C]

Relative humidity [%]

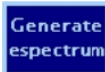


Figure A.7: HTML form to introduce input parameters (<http://fvred1.ctima.uma.es>)

The URI for accessing the software is:

<http://fvred1.ctima.uma.es>.

The developed web application allows us to generate the solar spectrum for any location. The user should provide certain meteorological measurements (horizontal global irradiance, air temperature, wind speed and relative humidity).

With the input information provided by the user, we first load and execute the model to select the best cluster in order to obtain the shape of the spectrum. We then use a second model to estimate the normalization factor. By using the normalization factor, the predicted spectrum is obtained (see Figure A.8). It is possible to download a file with the predicted spectra.

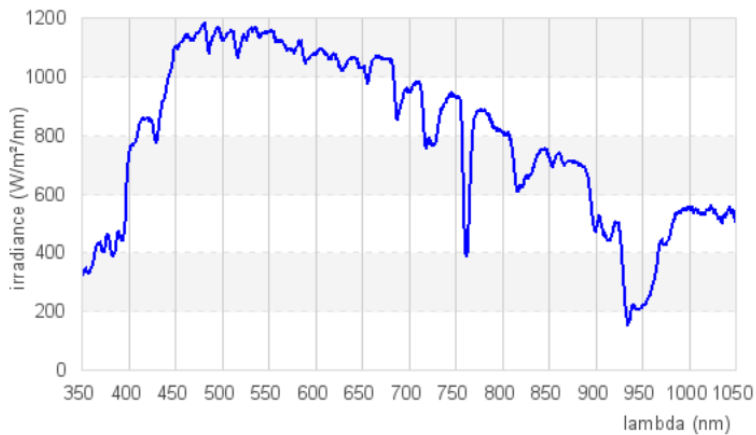


Figure A.8: Predicted spectrum using input data of Figure A.7 (<http://fvred1.ctima.uma.es>)