Alexandria Engineering Journal 79 (2023) 21–33

Contents lists available at ScienceDirect

Alexandria Engineering Journal

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

ORIGINAL ARTICLE

Neuro-fuzzy systems for daily solar irradiance classification and PV efficiency forecasting

Andrés Gersnoviez *, Juan C. Gámez-Granados, Marta Cabrera-Fernández, Isabel Santiago, Eduardo Cañete-Carmona, María Brox

Department of Electronic and Computer Engineering, Escuela Politécnica Superior, Universidad de Córdoba, Córdoba 14071, Spain

ARTICLE INFO

Keywords:
Irradiance daily profiles
Neuro-fuzzy system
Fuzzy classifier
Forecasting system
Photovoltaic installation efficiency

ABSTRACT

Considering the impact of photovoltaic installations and the fact that their performance depends on the type of day, this paper presents a classifier that makes use of fuzzy logic to classify daily irradiance profiles as a human would do. To do this, the system must be linguistically interpretable, so the classifier must be simple enough, but without losing accuracy. This is why the article combines the use of data mining and supervised learning algorithms to obtain an initial system and then exploits simplification techniques such as the concept of fuzzy classifiers with incomplete rule bases, as well as fuzzy tabular simplification of rules to obtain a compact and simple final system. The classifier obtained handles the ambiguity presented by the daily irradiance profiles with precision. Once the system has been obtained, a large number of days in southern Spain are classified, analysing the performances of a photovoltaic plant obtained in each of the classes. Then, a neuro-fuzzy system is designed to predict the performance of the photovoltaic installation, considering the type of day, the maximum ambient temperature reached during the day, and the degradation of the installation over time, proving its usefulness in alerting about anomalous behaviour of the system.

1. Introduction

The demand for electricity is increasing due to the growth of industrialization. Much of the electricity generated for lighting, heating, and cooling comes from fossil fuels, particularly coal and oil. The big disadvantage of using fossil fuels is that this has an environmental impact due to the carbon dioxide and mercury emissions produced by them, which is damaging to global warming of the planet.

In order to reduce this environmental impact one solution is to make use of renewable energies. Within renewable energies, solar energy is one of the cleanest and most abundant renewable resources available. In particular, solar photovoltaic (PV) energy uses solar radiation to produce electricity through the photoelectric effect using a PV cell. One of its advantages is that, as it is modular, it can be built from huge PV plants to small roof panels, and it represents the best economical and technological choice to electrify remote areas where the installation of electricity is expensive or difficult to reach. Solar PV energy is silent, easy to maintain and handle, and does not generate residues or polluting emissions. However, it is a variable energy, because it depends on meteorological factors that cannot be controlled (including the climate itself and air pollution) [1]. The solar plant output is generally proportional to solar irradiance so variations in irradiance cause fluctuations in plant output. Thus, it is very important to know exactly the solar irradiation resource that will be available at a location in order to design the appropriate PV system sizing (PV power, storage capacity). Therefore, given the increasing popularity of solar power generation worldwide, it is very interesting to develop a classifier that makes it possible to catalogue the solar irradiance profiles of an area collected by a PV installation. The classification of days allows to know the frequency at which each type of day occurs.

Classification of solar irradiance data to forecast the power output of a solar plant has been investigated in the literature [2]. Many studies have discussed the problem of typical day classification. These works differ by the parameters used as the criterion for the classification, the number of groups classified, and irradiance data were used with different temporal resolutions.

* Corresponding author.
E-mail addresses: andresgm@uco.es (A. Gersnoviez), jcgamez@uco.es (J.C. Gámez-Granados), martacabrerafdez@gmail.com (M. Cabrera-Fernández), el1sachi@uco.es (I. Santiago), ecanete@uco.es (E. Cañete-Carmona), mbrox@uco.es (M. Brox).

https://doi.org/10.1016/j.aej.2023.07.072
Received 27 October 2022; Received in revised form 10 May 2023; Accepted 25 July 2023

In [3] sky conditions are categorized into three groups (overcast, partly cloudy, and clear) using climatic data including cloud cover, sunshine hour, and solar radiation. [4] uses hourly horizontal and inclined solar irradiation collected every minute. The study allows showing that the four years of daily data can be clustered into three "typical" day classes (clear sky, overcast sky, cloudy sky). The method for classification followed by them uses Ward's classification process and discriminant analyses to establish the clustering procedure. In order to evaluate the performance of the PV system, [5] proposes a classification of days of three groups (clear sky, partially clouded, and opaque) using as parameters the monthly average of the global horizontal irradiation, the temperature, and the wind speed. [6] performs an analysis of daylight and solar radiation data classifying three sky conditions: clear sky, intermediate sky, and overcast sky. In order to develop this classification, two methods are used: cloud ratio method and sunshine duration where cloud ratio is defined as the proportion of the diffuse irradiance to the global irradiance. In this same line, the analysis proposed by [7] classified daylight data in Hong Kong into three sky conditions (clear, intermediate, and overcast) using two methods, that is, sunshine duration and cloud ratio methods.

There are works that use variability index as a parameter for their classification [8–11]. [8] describes a new and novel metric for quantifying variability in irradiance at solar PV sites. This metric measures the amount of variability in irradiance relative to the variability of a clear sky reference. Using this metric along with a daily clearness index (ratio of solar energy measured on a given surface to the theoretical maximum energy on that same surface during a clear sky day) a classification scheme that distinguishes four types of irradiance days, clear, overcast, mixed (clear/overcast) and highly variable, is presented. [9] uses combinations of the daily clearness index and the variability index in order to classify days with five categories: high variability, moderate variability, mild variability, clear, and overcast days. Following the same line, [10] uses the variability index to quantify the daily variability of solar irradiance using collected data from four stations in Australia. Variability is also used as a representative signal of daily behaviour in [11].

Other works perform an estimation of the fractal dimension of daily solar irradiance [12–15]. The fractal dimension is an important parameter that, for solar radiation, measures signal shape irregularity, describing the fluctuations of the phenomenon resulting from weather conditions. In this line [12] presents a model based on the Minkowski-Bouligand dimension. This work uses the fractal index and the clearness index to classify daily irradiances resulting in three classes: clear sky, partly cloudy sky, and completely cloudy sky. Another similar paper that also uses the fractal and clearness indexes as classification criteria is [13] where three classes are proposed: clear sky, partially covered sky, and overcast sky. [14] also proposes a classification method of daily solar irradiances based on fractals. Finally, [15] presents a method for the classification of daily irradiations by using four parameters: direct normal irradiation fraction, fractal dimension, and a variation coefficient of the time passed between two consecutive decreases of irradiation under two given thresholds. With this method four classes have been obtained: high daily energy (no irregularities on the daily curve), moderate daily energy (sky smoothly perturbed), moderate daily energy (sky very perturbed), and low daily energy (sky cloudy, all day).

On the other hand, [16] uses a spectral analysis of the wavelet transform by using 10 seconds of solar irradiation data and identifying three different classes (overcast, partly cloudy, and sunny). Another method that uses the transformation of temporal datasets is proposed in [17]. This paper classifies solar radiation daily patterns into four classes: clear sky, intermittent clear sky, completely cloud sky, and intermittent cloud sky. In order to perform this classification, a pair of indices are used: area ratio (A_r) and intermittency (I), which is calculated using the spectral power of clear sky irradiance. Finally, [18] proposes the use of wavelet transform to extract clear observation days from solar irradiance measurements collected in four years of data on global solar irradiance.

On the other hand, the use of clustering algorithms has been employed by various works to obtain groups of irradiance data [19–21]. In this line, [19] proposes a new method of classification of solar irradiance curves based on Mathematical Morphology techniques. This method focuses on the shape of the curves to form groups where cloud transitions and variability in direct radiation are taken into account. Another work with clusters is presented in [20]. This work shows a normalization of the solar irradiance data in order to make the time periods and solar irradiance comparable. In order to express all the patterns with the same number of points within the normalized time interval, interpolation is also performed. The patterns are then grouped using clustering, identifying four day types. Following these ideas, [21] presents a comparison of six clustering algorithms (K-means, hierarchical WMV, fuzzy c-means, self-organizing maps, ant-colony, and bat) from different clustering categories in order to investigate the appropriate method for establishing the grouping process of power patterns.

In another line, [22] presents a classification of days in tropical climates where the variation of solar radiation fluctuates greatly, due to the large amount of clouds. In order to characterize this fluctuation, the variations of daily solar radiation are summarized by a histogram where the histogram is an estimation of the measured daily solar radiation distributions. These histograms are classified by estimating a finite mixture of Dirichlet distribution obtaining four classes of distributions corresponding to four types of days. Another work that uses probabilistic representations is [23] where the measured solar energy data are classified into three groups using Gaussian Mixture Models.

There are works with a high number of classes of days [24–32]. [24] applies a classical maximum-likelihood method for clustering global and diffuse solar radiation data classifying sky conditions in nine classes. [25] is another example of a work that classifies a high number of groups. Specifically, seven sky conditions were found. The method employed uses the values of insolation indices to classify sky conditions from overcast to clear. In the same way, [26] presents a new methodology to classify ten types of days using clustering techniques according to the state of the sky based on three features: variability, energy, and time distribution of this energy. Clustering analysis is applied to thirteen years of 10-min measurements collected. Another work that uses ten classes corresponding to different sky conditions is proposed in [27]. [28] uses the monthly average of the global horizontal irradiation, the temperature, and the wind speed combined with the partial vapor pressure, the wind direction, and the sky nebulosity to obtain ten classes of days. Global horizontal irradiation recorded with an acquisition time step of 10 minutes is used in [29] to obtain nine day periods by decomposing each studied day in three periods. [30] presents four groups (nine sub-classes) for "day" periods and equally four groups (nine sub-classes) for "night" periods using horizontal irradiation, air temperature, wind speed, humidity, and nebulosity as main parameters to study the transitions between diurnal and nocturnal periods. Another related work is presented by [31] where a classification matrix for daily irradiance types is described. This matrix classifies the daily irradiance by the total amount of irradiance and irradiance fluctuations, resulting in nine different daily irradiance types. Finally, in [32] a day-type classifier based on the k-means algorithm is presented, dividing the days into 9 different classes, using the total irradiation values as the dominant parameter and the sum of irradiance variations as the second parameter. The results obtained by this classifier are very interesting, but there is a problem since, at first sight, it is not possible to differentiate the profiles obtained by the different classes. In other words, a human would not be able to distinguish one class from another without the help of the classifier system.

In relation to this last point and in order to solve the problem of the uncertainty and ambiguity involved in distinguishing one type of day from another, a powerful tool that can be used is fuzzy logic [33]. Fuzzy logic provides a mathematical framework to deal with the imprecision typical of the human reasoning system and has been applied in many disciplines, being one of them its use in renewable energies

(a) (b)

Fig. 1. PV installation: (a) BP-3165 solar modules; (b) SMC-100 inverters.

[34]. Fuzzy logic allows structuring the knowledge of a system through a set of symbolic "if-then" rules that use natural language terms to represent information. One line that many researchers have followed is to integrate the fuzzy systems paradigm with neural networks in so-called neuro-fuzzy systems in order to combine the advantages of both. A neuro-fuzzy system is basically a fuzzy system (hence with structured knowledge) that uses learning methods borrowed from neural systems to adjust its parameters [35].

In addition to the problem of typical day's classification, another interesting idea would be to forecast the performance of solar systems. Due to uncertainty involved in weather conditions, fuzzy logic has been proposed by researchers as a good tool for the performance prediction of solar PV systems. Numerous prediction models have been developed using artificial intelligence techniques [36,37]. Some works such as these proposed in [38,39] use neural networks in order to forecast solar power generation. In [38], in addition to a neural network-based model, an adaptive neuro-fuzzy interface-based model is also developed. [39] includes a comparison between two methods proposed in the work (one based on fuzzy logic and the other on neural networks) for forecasting solar power generation. Other works use fuzzy logic for the design of their prediction models [40–43]. Specifically, a type 2 fuzzy prediction system is presented in [41] and [43]. In [43] a type-2 Takagi-Sugeno-Kang fuzzy system is developed and root mean square errors (RMSE) of forecasts in the different seasons of the year (spring, summer, autumn, and winter) are presented. The previous prediction models proposed in the literature are limited because the degradation of solar systems through time, that affects the performance of the PV systems, is not considered.

Therefore, this paper proposes the development of a novel fuzzy classifier of days with a high number of classes using daily irradiance measurements to extract the rule base. The aim is to achieve a simple and linguistically interpretable system that can be understood by non-experts, but without sacrificing accurate and robust behaviour. In order to achieve a precise behaviour, data mining and supervised learning techniques will be used; in order to be linguistically interpretable, on the one hand as few parameters as necessary will be sought, and on the other hand the use of simplification techniques will be exploited to achieve the simplest possible system. In addition, with this classifier, a prediction fuzzy model which allows forecasting the performance of solar panels has been developed. This proposed prediction model provides the novelty of considering the degradation of the performance of solar systems in time.

The structure of the paper is as follows. Section 2 will focus on describing the climatic characteristics of the city in which the PV installation that is going to be studied is located, as well as the instruments that the plant possesses, and which have been used to record the data used in this work. Also, the parameters to determine the performance of PV plant components are described; Section 3 will specify the types of days that the classifier will take into account, as well as the parameters that determine them; Section 4 will show how the fuzzy classifier has been designed, as well as the results obtained from its behaviour and a discussion about them; Section 5 will focus on the design of the forecasting system for the performance of the PV installation, showing the results obtained from it, as well as a comparison with other systems developed with other learning algorithms; finally, the paper finishes with Section 6, which is dedicated to the conclusions.

Table 1
Characteristics of the PV system.

Area of each panel	1.573*790 mm^2
Current at maximum power (I_{max})	4.7 A
NOCT	47 °C ±2 °C
Number of inverters	3
Number of panels per inverter	36 (12*3 modules)
Number of panels in plant	108
Number of cells/panel	72
Open-circuit voltage (V_{oc})	44.2 V
Power of panels per inverter	5.940 W (36 panels * 165 Wp)
Power temperature coefficient	-(0.5 ±0.005)%K
PV cell surface	125*125 mm^2
Short-circuit current (I_{sc})	5.2 A
Total power	17.920 W (3 inverters * P_{inv})
Useful panel area	1.125 m^2
Voltage at maximum power (V_{max})	35.2 V

2. Characteristics of the local climate and measuring instruments

The climatological characteristics of the city of Córdoba, located in southern Spain, make it very suitable for obtaining solar energy, since it is a city with a Mediterranean climate, mostly arid and cloudless, with little rainfall. In particular, winters in Córdoba are soft, without reaching extreme temperatures, except on rare occasions when frosts may occur, with January being the coldest month, where the average minimum temperature is 4 °C, rarely reaching -1 °C. On the other hand, rainfall is concentrated in the months of spring and autumn, reaching an average of 600 mm per year, so they are not abundant, with a lot of irregularity from one year to another. And finally, summers are extremely hot, being July the hottest month, reaching temperatures of 40 °C during the day and 18 °C at night, making it one of the hottest cities in Europe, and occasionally appearing in the 50 hottest cities in the world ranking.

Concerning the characteristics of the rooftop PV installation, it has a peak power of 17.82 kW, is formed by BP solar modules (model BP-3165), which are made by polycrystalline silicon and have a peak power of 165 Wp. These modules are mounted on a 30° open frame, 18° south offset (Fig. 1(a)). The modules are connected in strings of twelve units connected in series, and three strings of this type are connected in parallel directly to each of the three inverters in the system (Fig. 1(b)). The inverters are from SMA (model SMC-100). Tables 1 and 2 show the characteristics of the PV system and the inverters, respectively.

The data acquisition system is located in the inverters of the PV plant. The irradiance data are collected by a device called Sunny Sensor-Box, from the SMA company, where the irradiance sensor is connected, together with two temperature sensors to register the values of the ambient and module temperatures, which are listed in Table 2.

Table 2
Characteristics of the inverters of the plant.

European performance	95.20%
Max. AC power	5500 W
Max. DC power	5750 W
Max. DC voltage	600 V
Max. input current	26 A
Max. nominal AC power	5000 W
Max. output current	26 A
Max. performance	96.10%
Number of MPPT	1
PV voltage range, MPP	246-480 V

- Amorphous ASI PV cell, calibrated for irradiance acquisition, with a measurement range from 0 to 1500 W/m^2, an accuracy of ±8%, and a resolution of 1 W/m^2.
- PT-100M temperature sensor, with a range of -20 to +110 °C, an accuracy of ±0.5 °C, and a resolution of 0.1 °C, used to measure the temperature of the modules.
- PT-100M-NR sensor, same characteristics as the previous one, but with an accuracy of ±0.7 °C, used to measure the ambient temperature.

The SensorBox and the inverter are connected via RS485 to a device called SunnyWebbox from SMA Company. It is a device with a central communications unit that continuously collects all the data from the inverter and the SensorBox and is a multifunctional, energy-saving data logger. The measurement data are transmitted through a GSM modem from remote locations where there is no telephone or ADSL connection, so they are read directly from a computer connected to the network, consuming between 4 and 12 W of power.

With these instruments, the measurements were taken every 5 minutes (i.e. 288 measurements per day) for 3322 days (slightly more than 9 years) in the city of Córdoba. Specifically, these measures were carried out between January 1, 2011, and February 4, 2020. Of all these days, only 3158 were finally taken into account after the correction of errors (for example, days when the system stopped measuring for several hours).

It will be interesting to see what kind of performance is obtained depending on the type of day. For this purpose, the efficiency values to be taken into account in this work are those related to PV modules and inverters. In addition to these two values, the *Performance Ratio* (*PR*), which is a dimensionless parameter corresponding to the ratio of the energy generated and delivered to the grid, compared to the value of the energy generated by the same system without losses, will also be taken into account.

For the calculation of these parameters, the data provided by the inverter, corresponding to the inverter input current and voltage, and the inverter output power, have been used.

To calculate the efficiency of the PV modules, the following expression has been considered [44]:

$$\eta_G = \frac{I_{dc} \cdot V_{dc} \cdot t}{E_G \cdot A_G} \cdot 100 \tag{1}$$

where I_{dc} is the direct current generated in the modules and supplied to the inverter; V_{dc} is the voltage at the inverter input from the modules; t is the time of the monitoring interval; E_G is the in-plane irradiance of the PV panels; and A_G is the area of the PV panels associated with each inverter.

Additionally, the following expression is considered for the calculation of the inverter efficiency:

$$\eta_{inv} = \frac{P_{ac}}{I_{dc} \cdot V_{dc} \cdot t} \cdot 100 \tag{2}$$

where P_{ac} is the output power of the inverter.

Finally, for the calculation of *PR*, the following applies:

$$PR = \frac{P_{ac} \cdot t \cdot G^*}{P_0 \cdot E_G} \tag{3}$$

where $G^* = 1000$ W/m^2 is the irradiance at standard STC conditions (AM1.5, 1 kW/m^2, 25 °C); and P_0 is the nominal power at standard STC conditions.

3. Types of days according to their daily irradiance shape

One of the objectives of this work is to develop a fuzzy system that is able to classify days into different types of profiles depending on their daily irradiance values in the same way as a human would do it. In this work the days will be classified into four macroclasses, which are *Sunny, Partly Sunny, Partly Cloudy* and *Cloudy*. *Sunny* days are those days with no clouds and their irradiance profiles have a gaussian bell shape. *Partly Sunny* days are those with small distortions in the bell due to some cloud passage. *Partly Cloudy* days are those with much more pronounced distortions than *Partly Sunny* days, caused by increased cloud passage. Finally, *Cloudy* days are those in which the sky is completely covered with clouds, and therefore no bell shape is visible. Each of these four macroclasses will be further divided into three subclasses, *High, Medium* and *Low*, according to the level of total irradiation received on the day. This is because a sunny summer day is not the same as a sunny winter day, as the hours of daylight received vary from one to the other. Thus, the twelve total classes will be: *Sunny High, Sunny Medium, Sunny Low, Partly Sunny High, Partly Sunny Medium, Partly Sunny Low, Partly Cloudy High, Partly Cloudy Medium, Partly Cloudy Low, Cloudy High, Cloudy Medium* and *Cloudy Low*. Examples of each of these 12 classes are shown in Fig. 2.

Once the daily irradiance profile classes have been selected, the next step is to specify which values determine that a day belongs to one class or another. In this sense, as the aim is to achieve a simple system, the number of parameters should be as small as possible and, taking into account the large amount of data to be processed, those with the lowest computational cost should also be chosen.

The irradiance values used in this work are global tilted irradiance (GTI), which are measured in the plane of the PV modules. After numerous studies, it was concluded that the value that determines which macroclass a profile belongs to is the sum of the variations of the instantaneous irradiance values (called *VAR* in this work). Specifically, taking into account that 288 measurements are taken each day, the *VAR* calculation expression would be:

$$VAR = \sum_{i=1}^{287} |GTI_i - GTI_{i-1}| \tag{4}$$

where GTI_i is the i-th measure of the GTI value for the day under analysis.

The initial idea was to use the total length of the curve as a parameter (which, in reality, is a piecewise-linear function), instead of *VAR*, but, taking into account that the length of the segment of the piece i, l_i would be:

$$l_i = \sqrt{(GTI_i - GTI_{i-1})^2 + (t_i - t_{i-1})^2} \tag{5}$$

and also, that the space of time between measurements is constant, the previous expression becomes:

$$l_i = \sqrt{(GTI_i - GTI_{i-1})^2 + const} \tag{6}$$

Therefore, the information that *VAR* gives us in this respect is equivalent to the length of the curve in this case, but with a much smaller computational weight.

Several profiles were taken that clearly belonged to each macroclass and it was found that the *VAR* values presented by each macroclass belonged to non-overlapping intervals, hence the *VAR* value was taken to determine to which macroclass a given profile belongs.

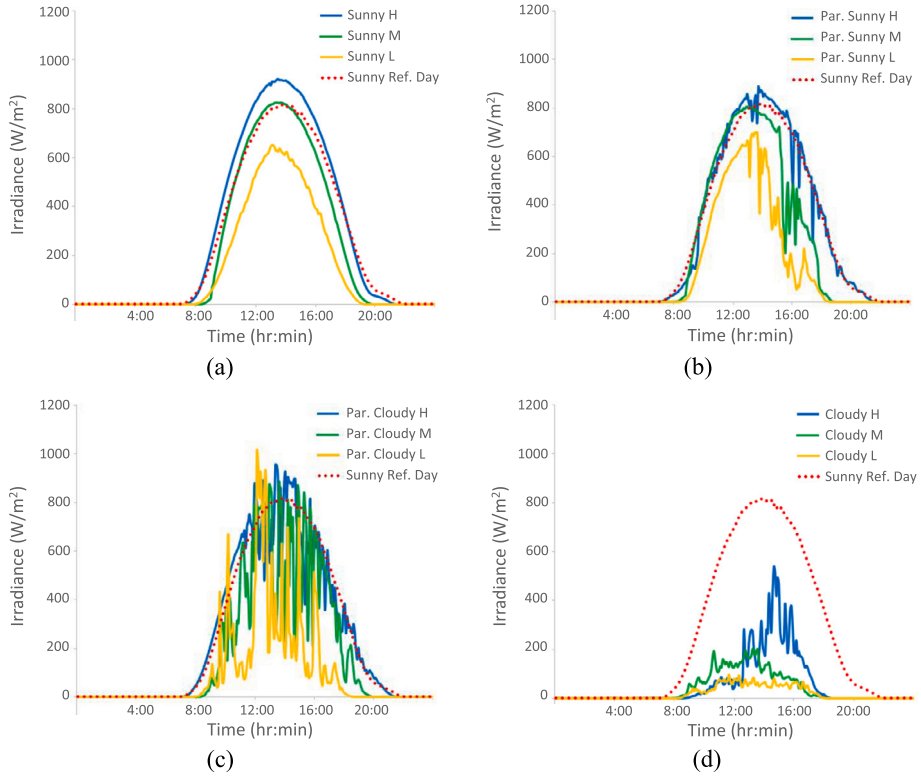

Fig. 2. Examples of profiles of the different types of days: (a) *Sunny* type days; (b) *Partly Sunny* type days; (c) *Partly Cloudy* type days; (d) *Cloudy* type days.

In addition, it was found that the different profiles grouped in each macroclass differed considerably in their total irradiance values, hence it was necessary to make a distinction between these types of days, generating the different subclasses. To calculate a value equivalent to the value of the total irradiance received on the day, but with a low calculation weight, it was decided to sum all the instantaneous irradiance values of that day (called *SUM* in this work). Specifically, the expression of *SUM* would be:

$$SUM = \sum_{i=0}^{287} GTI_i \tag{7}$$

Specifically, the interval of values of *SUM* for each macroclass can be determined, and each of these intervals can be divided into three subintervals: *High, Medium* and *Low*.

Finally, and with these assumptions, the fuzzy classifier can be developed.

4. Neuro-fuzzy classifier design

The reasons for choosing a fuzzy classifier are diverse. Firstly, one of the intrinsic characteristics of fuzzy systems is the use of concepts with linguistic interpretability. This factor means that non-experts can understand how the classifier works. That is, anyone can understand what a *sunny* or *partly cloudy* day is, then knowing what kind of day it is going to be, it is possible to determine what efficiency can be achieved with the PV installation.

On the other hand, Fig. 2 shows day profiles that clearly belong to one class or the other. However, there are day profiles that are halfway between one class and other and it is not entirely clear, at a first glance, to which class they may belong. This makes the use of fuzzy systems more than appropriate, due to the facility they have to handle uncertainty and ambiguity.

Among the different conditions reported in the literature for a system to be linguistically interpretable, the number of rules must be moderate and the membership functions must be distinguishable and cover the entire universe of discourse of the variables [45]. In turn, when designing the system, fuzzy grid-type systems will be chosen, because they facilitate its linguistic interpretability, as long as the number of rules is reduced.

However, as explained in [46] and [47], if a fuzzy grid-type classifier is chosen and the rule base is complete, then the system has too much information to allow ambiguity. That is why the rule base of the classifier must be incomplete, leaving gaps between the rules defining the 4 macroclasses (*Sunny, Partly Sunny, Partly Cloudy* and *Cloudy*) so that the transition between them is gradual and the behaviour of the system is truly *fuzzy*. In a non-fuzzy system with an incomplete rule base, the system may find itself in a situation where it does not know what output value to give. However, in a fuzzy system, the empty areas of the rule base are occupied by neighbouring rules, since their degree of activation is non-zero, as explained in [48], [46] and [47].

Therefore, the constraints imposed on the system are that it should be a fuzzy grid-type classifier, with a moderate number of rules and distinguishable membership functions covering the entire universe of discourse, to ensure that it is linguistically interpretable; and, on the other hand, that the rule base should be incomplete to ensure that the transition between macroclasses is gradual and fuzzy, to deal better with uncertain cases.

For the development of the classifier, days that clearly belonged to a particular macroclass were found to have *VAR* values in non-overlapping intervals with the other macroclasses. Therefore, the *VAR* values of these four intervals were separated and the *SUM* values in each of them were analysed. Knowing the minimum and maximum values of *SUM* in each interval, it was divided into three equal ranges to establish the subclasses of *Low, Medium* and *High* for each macroclass.

Once a training file has been created for which, to certain input values of *VAR* and *SUM*, an output day class is assigned, an identification algorithm is used to create the system. The algorithm chosen for this task is the Wang-Mendel algorithm [49] and, to tune the behaviour of the system obtained, the Levenberg-Marquardt algorithm (a second-order training algorithm that does not compute the Hessian ma-

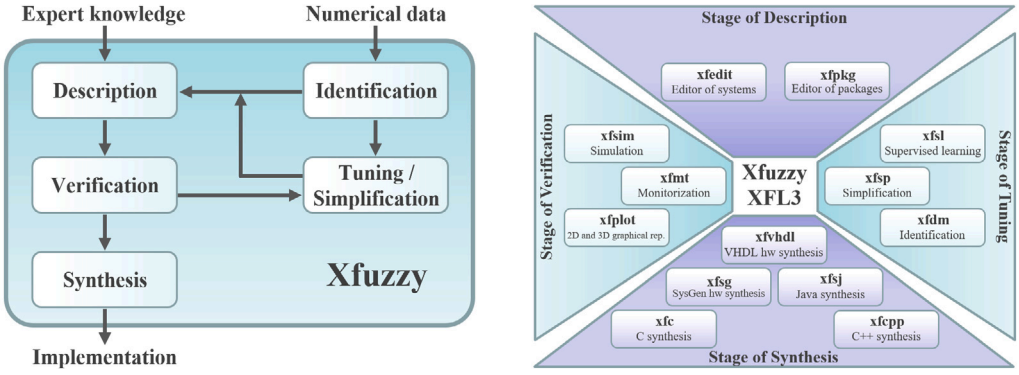

Fig. 3. (a) Design methodology flow followed by Xfuzzy; (b) Xfuzzy set of CAD tools, grouped according to the stage of design they relate to.

trix [50]) is used. For this purpose, the Xfuzzy environment [51] has been used.

The Xfuzzy environment, developed at the Instituto de Microelectrónica de Sevilla, allows the design of complex fuzzy systems characterised by being extensive, expressive and modular. The design flow that can be followed with Xfuzzy is shown in Fig. 3(a). Fig. 3(b) illustrates the set of CAD tools in the environment, grouped according to the design stage to which they relate. These tools cover all stages of the fuzzy system design process, from its initial description to its final implementation. The link between them is the use of a common specification language, *XFL3*.

The first step in the design is to use the Xfuzzy identification tool, *xfdm*, choosing the Wang-Mendel algorithm. The chosen defuzzification method is the one that outputs the consequent of the most activated rule (which we will call *MaxLabel* in this paper). Taking into account that the resulting rule base will have sufficiently large empty areas, the membership functions for the inputs must be Gaussian, as explained in [47]. This ensures that there are no undefined zones, since even if the degree of activation of a rule is very small, it is sufficient to dominate over smaller values and provide its category as output value.

Once the fuzzy classifier is obtained, the Levenberg-Marquardt algorithm has been applied with the help of the Xfuzzy supervised learning tool, *xfsl*, to tune the behaviour of the system. Finally, rules were manually deleted to ensure that there were empty areas between the sets of rules belonging to the different macroclasses.

After several tests, from 12 membership functions to describe the inputs (thus forming a 12x12 grid partition), no significant improvement in the behaviour of the system is achieved, and a higher number of functions starts to negatively affect the linguistic interpretability. The final Neuro-Fuzzy Classifier System (NFCS) obtained with 68 rules is shown in Fig. 4 and 5.

Although the number of rules is not too high, linguistic interpretability may be lost due to their number. Therefore, one option to improve the linguistic interpretability of the system is to simplify its rule base. Specifically, if the Fuzzy Tabular Simplification algorithm explained in [52] is used, being an extension of the Quine-McCluskey algorithm from Boolean logic to fuzzy logic, and integrated into the Xfuzzy simplification tool *xfsp*, the rules for each consequent can be automatically grouped, obtaining the following rule base composed of only 12 rules:

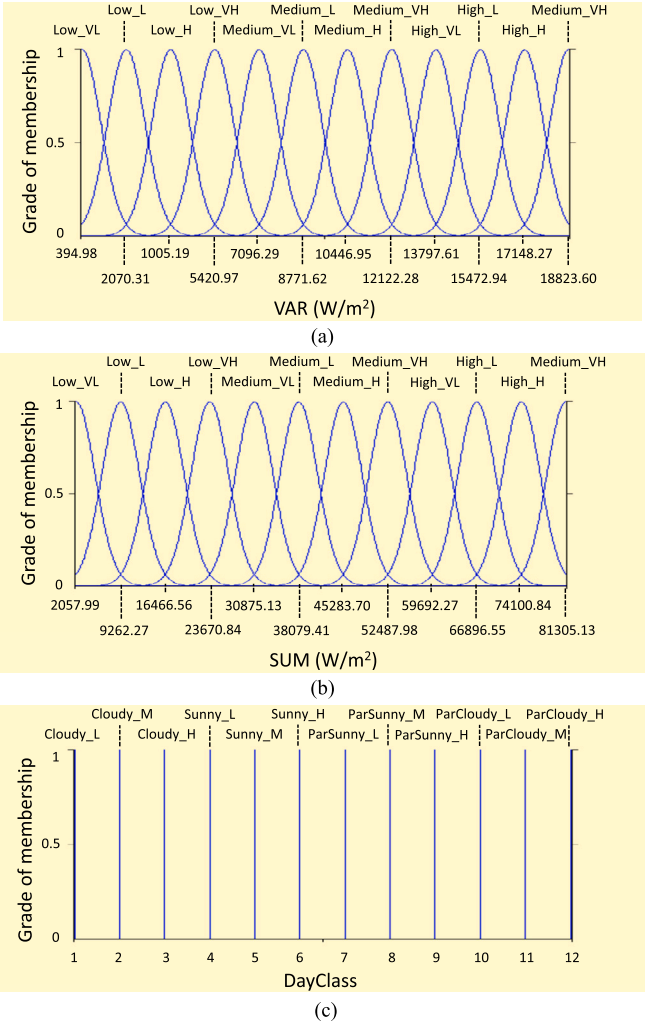

Fig. 4. (a) Membership functions for the *VAR* input; (b) Membership functions for the *SUM* input; (c) Membership functions for the *DayClass* output.

1. *If ((VAR is ≥ Low_VL and VAR is ≤ Low_L) and (SUM is Low_VL)), then DayClass is Cloudy_L;*
2. *If ((VAR is ≥ Low_VL and VAR is ≤ Low_L) and (SUM is Low_L)), then DayClass is Cloudy_M;*
3. *If ((VAR is ≥ Low_VL and VAR is ≤ Low_L) and (SUM is ≥ Low_H and SUM is ≤ Low_VH)), then DayClass is Cloudy_H;*
4. *If ((VAR is Low_VL) and (SUM is ≥ Medium_H and SUM is ≤ Medium_VH)), then DayClass is Sunny_L;*
5. *If ((VAR is Low_VL) and (SUM is ≥ High_VL and SUM is ≤ High_L)), then DayClass is Sunny_M;*
6. *If ((VAR is Low_VL) and (SUM is ≥ High_H and SUM is ≤ High_VH)), then DayClass is Sunny_H;*
7. *If ((VAR is Low_H) and (SUM is ≥ Medium_H and SUM is ≤ Medium_VH), then DayClass is ParSunny_L;*
8. *If ((VAR is Low_H) and (SUM is ≥ High_VL and SUM is ≤ High_L)), then DayClass is ParSunny_M;*

VAR \ SUM	Low_VL	Low_L	Low_H	Low_VH	Medium_VL	Medium_L	Medium_H	Medium_VH	High_VL	High_L	High_H	High_VH
Low_VL	Cloudy_L	Cloudy_M	Cloudy_H	Cloudy_H			Sunny_L	Sunny_L	Sunny_M	Sunny_M	Sunny_H	Sunny_H
Low_L	Cloudy_L	Cloudy_M	Cloudy_H	Cloudy_H								
Low_H							ParSunny_L	ParSunny_L	ParSunny_M	ParSunny_M	ParSunny_H	ParSunny_H
Low_VH												
Medium_VL							ParCloudy_L	ParCloudy_M	ParCloudy_M	ParCloudy_H	ParCloudy_H	ParCloudy_H
Medium_L							ParCloudy_L	ParCloudy_M	ParCloudy_M	ParCloudy_H	ParCloudy_H	ParCloudy_H
Medium_H							ParCloudy_L	ParCloudy_M	ParCloudy_M	ParCloudy_H	ParCloudy_H	ParCloudy_H
Medium_VH							ParCloudy_L	ParCloudy_M	ParCloudy_M	ParCloudy_H	ParCloudy_H	ParCloudy_H
High_VL							ParCloudy_L	ParCloudy_M	ParCloudy_M	ParCloudy_H	ParCloudy_H	ParCloudy_H
High_L							ParCloudy_L	ParCloudy_M	ParCloudy_M	ParCloudy_H	ParCloudy_H	ParCloudy_H
High_H							ParCloudy_L	ParCloudy_M	ParCloudy_M	ParCloudy_H	ParCloudy_H	ParCloudy_H
High_VH							ParCloudy_L	ParCloudy_M	ParCloudy_M	ParCloudy_H	ParCloudy_H	ParCloudy_H

Fig. 5. Classifier system rule-base.

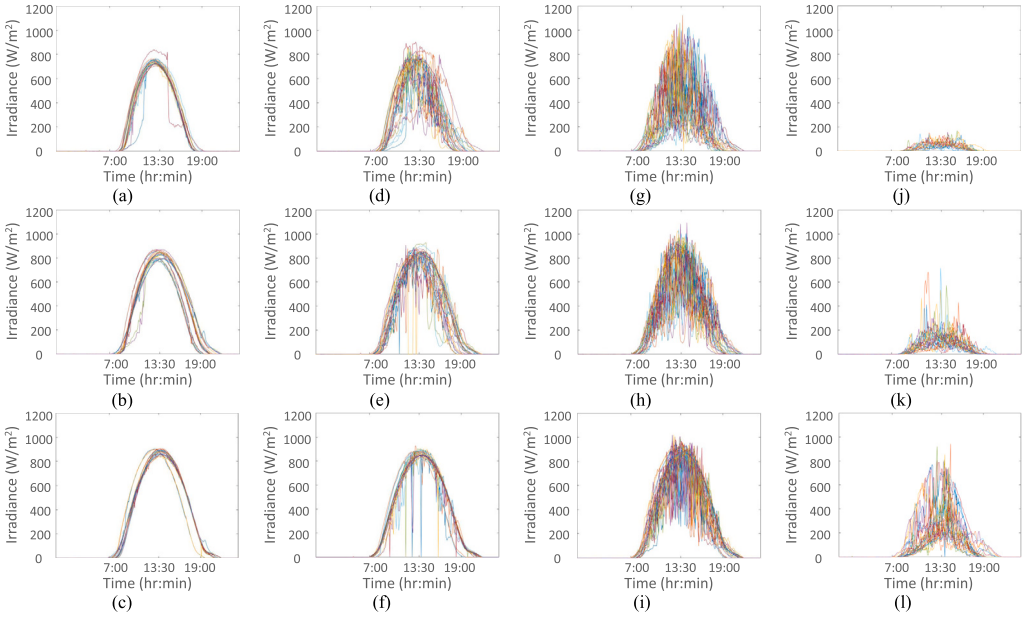

Fig. 6. Sample of daily irradiance profiles obtained by the NFCS for: (a) *Sunny Low*; (b) *Sunny Medium*; (c) *Sunny High*; (d) *Partly Sunny Low*; (e) *Partly Sunny Medium*; (f) *Partly Sunny High*; (g) *Partly Cloudy Low*; (h) *Partly Cloudy Medium*; (i) *Partly Cloudy High*; (j) *Cloudy Low*; (k) *Cloudy Medium*; (l) *Cloudy High*.

9. *If ((VAR is Low_H) and (SUM is ≥ High_H and SUM is ≤ High_VH)), then DayClass is ParSunny_H;*
10. *If ((VAR is ≥ Medium_VL and VAR is ≤ High_VH) and (SUM is Medium_H), then DayClass is ParCloudy_L;*
11. *If ((VAR is ≥ Medium_VL and VAR is ≤ High_VH) and (SUM is ≥ Medium_VH and SUM is ≤ High_VL)), then DayClass is ParCloudy_M;*
12. *If ((VAR is ≥ Medium_VL and VAR is ≤ High_VH) and (SUM is ≥ High_L and SUM is ≤ High_VH)), then DayClass is ParCloudy_H;*

As can be seen, the rule base is sufficiently compact and simple to maintain high linguistic interpretability.

4.1. Results and discussion of the classifier system

Once the system has been completed, its correct performance in classifying day profiles is tested. For this purpose, the Xfuzzy simulation tool, *xfsim*, is used to perform a closed-loop simulation. In this simulation, a plant described in Java has been created with which a file with the *VAR* and *SUM* values of the 3158 days analysed has been run through and the system has returned the corresponding day class for each of these days. The number of days per class obtained is shown in Table 3.

The results shown in the table are consistent with the type of weather in the city, with sunny days being the most numerous (with the *Sunny High* subclass being by far the most dominant, which was to be expected), followed by partly sunny days, and cloudy days being by far the least numerous.

Table 3
Number of days per class according to the NFCS.

Day class	Number of days per subclass	Number of days per macroclass
Cloudy Low	17	
Cloudy Medium	108	410
Cloudy High	285	
Sunny Low	224	
Sunny Medium	301	1171
Sunny High	646	
Partly Sunny Low	260	
Partly Sunny Medium	272	811
Partly Sunny High	279	
Partly Cloudy Low	345	
Partly Cloudy Medium	225	766
Partly Cloudy High	196	

However, what is really interesting is to check that the daily profiles adjust to what is shown in Fig. 2. For this purpose, Fig. 6 shows samples of different daily profiles assigned according to the NFCS.

As can be seen, the results obtained are highly satisfactory. In the figure it can be seen how, indeed, sunny days show bells with no noise; partly sunny days show bells with some noise; partly cloudy days show bells with a lot of noise; and, finally, cloudy days show no bell shape.

To compare this system with other authors is difficult, since, as could be seen in Section 1, there is no common agreement on the number of classes into which the daily irradiance profiles should be divided.

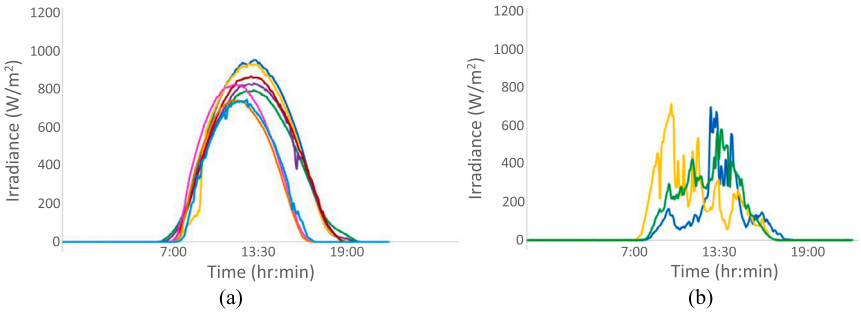

Fig. 7. Cases of extreme confusion of the Table 4: a) Days classified by AEMET as *Cloudy* and by NFCS as *Sunny*; b) Days classified by AEMET as *Sunny* and by NFCS as *Cloudy*.

Moreover, the results are highly dependent on the climatology of the location, varying greatly from one country to another.

However, there is one work with which it can be compared, and that is the one presented in [32], because they use the same database that has been used in this work (with the same PV system in the same city).

In [32] they use as dominant parameter the sum of all the instantaneous irradiance values, due to its importance in obtaining good performance values. However, not using *VAR* as the dominant parameter means not determining whether the profile shows a clean, low-noise or high-noise Gaussian bell. That is, they sacrifice an important property in this type of systems, which the NFCS largely achieves: that the profiles can be classified as a human would do it.

The profiles obtained by the classifier of [32] have interesting properties, but a human would be unable, given a profile, to determine which class it belongs to of those proposed by the authors. The classifier is completely necessary to determine to which class it belongs. However, given a profile, a human would be able to distinguish between the classes shown in Fig. 2 and 6, which is what the classifier described in this paper does.

This is important because the performances achieved are highly dependent on the type of day. And, if the aim is to know in advance the performances that are going to be achieved according to the predicted day, weather forecasts are based on reporting how clear the day is going to be, as well as the temperatures that are going to be reached.

Another way to keep checking the good performance of the NFCS is to corroborate what kind of day, according to the sky conditions, were the days analysed. To do this, we have contacted the State Meteorological Agency of the Spanish Government (AEMET), and we have asked them to provide us with data on the state of the sky in Cordoba on those days. The station where the agency takes measurements in Cordoba is 7.2 km away from the PV installation analysed in this article. In addition, for the historical sky state data records, the method followed by the agency is that a technician observes the sky at 7 a.m., 1 p.m., and 6 p.m., and notes the corresponding oktas, taking into account that the oktas values range from 0 to 8, with 0 being a sky completely clear, and 8 being a sky completely cloudy. To compare with the results of the developed NFCS, the mean value of the 3 oktas values for each day has been calculated and the interval between 0 and 8 has been divided into 4 sub-intervals, the first one being for *Sunny* days, the next one for *Partly Sunny*, the next one for *Partly Cloudy* and the last one for *Cloudy* days.

Removing the days with incomplete data provided by AEMET, 2523 days remain for comparison. Once the results of the NFCS have been compared with those of AEMET, the confusion matrix shown in Table 4 is obtained.

As can be seen in the Table 4, a typical good classification behaviour can be observed, having the highest values on the diagonal. As, in addition, the classes are ordered from one extreme case (completely clear sky) to another (completely cloudy sky), it can be seen that the values, as they move away from the diagonal, decrease. The reason for the discrepancies between the AEMET data and the NFCS results are not due to failures of the NFCS, but mostly due to differences in the way the data is collected and the location. On the one hand, the AEMET measurements are made by a human at three specific times of the day (and, between these times, there can be important variations that are not recorded and that can change the classification completely), while the measurements used by the NFCS are made by an automatic measurement system every 5 minutes throughout the day. On the other hand, although 7.2 km (which is the distance between the AEMET station in Cordoba and the PV installation studied) is not a large distance, it may be enough to vary the state of the sky from one point to another. To reinforce this argument, Fig. 7 shows the most extreme cases of confusion, such as the 8 days that AEMET classifies as *Cloudy* and the NFCS as *Sunny*, and the 3 days that AEMET classifies as *Sunny* and the NFCS as *Cloudy*. As can be seen, the NFCS correctly assigns the classes for these days. Therefore, a system such as the one developed in this article could be used to debug errors in the methodology used by AEMET.

Table 4
Confusion matrix of the data provided by AEMET and the NFCS results.

AEMET \ NFCS	Sunny	Partly Sunny	Partly Cloudy	Cloudy
Sunny	743	145	27	3
Partly Sunny	157	309	155	16
Partly Cloudy	17	139	289	59
Cloudy	8	48	178	230

In addition to the confusion matrix, different metrics can be used to check the good performance of the NFCS, such as *Precision* and F_1-*Score* [53], which combines the values of *Precision* and *Recall*. Specifically, *Precision* details the percentage of patterns labelled as a class that actually correspond to that class, while *Recall* details the proportion of patterns of a class that the model has been able to correctly label:

$$Precision = \frac{TP}{TP+FP}; \quad Recall = \frac{TP}{TP+FN} \tag{8}$$

where TP is the number of instances that the classifier has identified of a specific class correctly (true positives); FP the number of instances that the classifier has identified of that specific class incorrectly (false positives); and FN the number of instances that the classifier should have classified of that specific class and did not (false negatives).

Once the equations for calculating *Precision* and *Recall* have been defined, the F_1-*Score* is calculated as:

$$F_1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{9}$$

The *Precision* and F_1-*Score* results obtained for the *Sunny*, *Partly Sunny*, *Partly Cloudy* and *Cloudy* macroclasses of the NFCS are shown in the Table 5. To reinforce the good performance of our proposal, this table also shows the results obtained from two other systems based on the following algorithms: Adaptive Neuro-Fuzzy Inference System (ANFIS), which is a machine learning algorithm that uses a neural network together with fuzzy logic techniques to create a fuzzy inference system [54]; and C4.5, which is one of the most widely used machine learn-

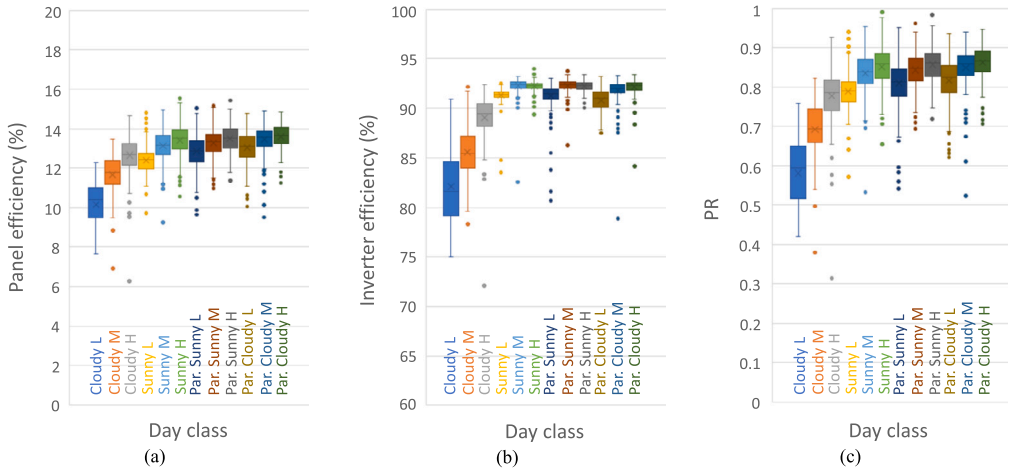

Fig. 8. (a) Panel efficiency according to day class; (b) Inverter efficiency according to day class; (c) *PR* according to day class.

Table 5
Comparison of performance assessment parameters for various approaches to the classification problem.

Class	ANFIS		C4.5		NFCS	
	Prec.	F_1-Sc.	Prec.	F_1-Sc.	Prec.	F_1-Sc.
Sunny	57.2%	67.3%	84.1%	77.8%	**80.3%**	**80.6%**
Partly Sunny	37.1%	37.5%	33.7%	47.4%	**48.2%**	**48.4%**
Partly Cloudy	37.6%	23.8%	45.3%	23.7%	**44.5%**	**50.1%**
Cloudy	30.1%	24.7%	83.3%	13.8%	**74.7%**	**59.6%**
Complete Syst.	46.7%	38.3%	51.1%	40.7%	**62.7%**	**59.7%**

ing algorithm in classification problems that uses relative information gain (entropy) to generate a decision tree [55]. The ANFIS system has been implemented using Matlab, while the C4.5 system has been implemented using WEKA software [56] [57]. For both, the default values defined in the implementation of the algorithms have been used.

Analysing the results shown in the Table 5, it can be seen that, of all the systems, the one that shows the best performance is the NFCS, being the one that presents the best F_1-*Score* values in all the fields, and also the one that shows the least differences between the *Precision* and F_1-*Score* values.

Once the NFCS has been checked and a class has been assigned to each day, the days can be grouped by class, and the values of the parameters η_G, η_{inv} and *PR* can be compared according to the day class, obtaining the results shown in Fig. 8. Also, the values obtained for maximum ambient temperature and maximum module temperature can be grouped according to the day classes, getting the results illustrated in Fig. 9.

By observing the different subclasses belonging to the same macroclass, it can be seen that the best-combined performance of modules and inverters, set by the value of *PR*, happens at higher values of total irradiance (i.e. of each macroclass, the best-performing subclass is the one with the descriptor *High*).

From Fig. 9, it can be observed that the *Cloudy* macroclass, closely followed by the *Sunny Low* subclass, are the ones with the lowest temperature values, corresponding to winter days (being especially logical in *Sunny Low*, where, even with clear skies and bell-shaped profiles, low irradiation values are obtained due to fewer hours of sunshine). At the other extreme, the subclasses with the highest number of hot days, from highest to lowest, are *Sunny High, Partly Sunny High, Sunny Medium, Partly Sunny Medium, Partly Cloudy High*. The latter is important for analysing the performance behaviours achieved by the different classes (Fig. 8), as will be discussed below.

It might have been thought that the best *PR* values would be achieved with those classes with the highest total irradiation and clear

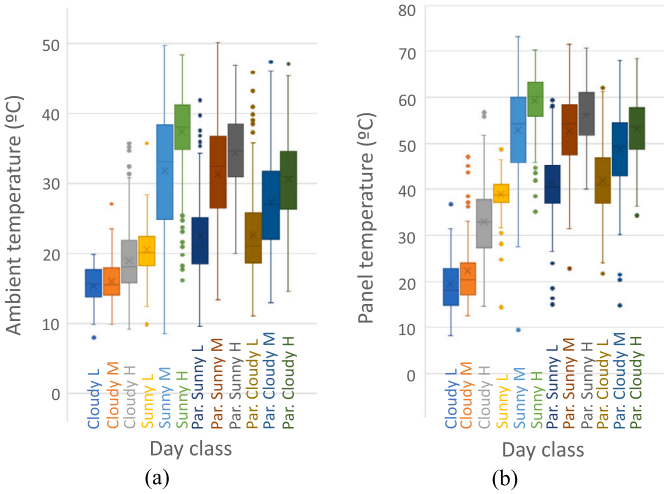

Fig. 9. (a) Daily maximum ambient temperature values according to day class; (b) Daily maximum temperature values of the panels according to day class.

skies (*Sunny High* and *Sunny Medium*), as clouds would deprive the panels of some of the irradiation received. However, analysing the results obtained from *PR*, although the values obtained are close, the classes with the best performance, from highest to lowest, are *Partly Cloudy High, Partly Sunny High* and *Sunny High* (Fig. 8(c)). Although at first glance the result may seem surprising, these results have a double justification.

The first can be deduced from the results in Fig. 8(a) and Fig. 9. It can be seen that the panels perform better on *Partly Cloudy High* days than on *Partly Sunny High* and *Sunny High* days. Fig. 9 shows that the *Partly Cloudy High* days are colder days, a factor to be taken into account since the overheating of the panels implies energy losses [58] (especially considering the extremely high temperatures reached in the city).

On the other hand, another phenomenon must be taken into account, and it is the so-called *Cloud Enhancement* or *Irradiance Enhancement*, studied by many authors [59–61] [32]. This phenomenon can be observed simply in Fig. 2 and in a more dramatic way in Fig. 6. When following the bell shape of clear days, when there is a drop in the irradiance value due to clouds, followed by a rise, the irradiance value does not return to the trajectory imposed by the bell but exceeds it. And these peaks rise more sharply if the variations are more drastic. These peaks in the irradiance values are due to reflection phenomena produced by the cloud edges surrounding the unobstructed solar disc.

Therefore, the irradiance lost by the clouds is then compensated by the over-irradiance produced by this phenomenon. It is also for this reason that, in the inverter performance results (Fig. 8(b)), the *Partly Cloudy High* and *Medium* class obtain similar results to *Sunny High* and *Medium*, since, due to over-irradiation, they obtain similar values of total irradiation received in the end.

Therefore, the positive results obtained on *Partly Cloudy High* days are a combination of the reduction of losses due to the cooling of the panels, together with the reception of total irradiation values similar to those of a clear day, due to the over-irradiation obtained by the *Cloud Enhancement* phenomenon.

5. Neuro-fuzzy forecasting system design

The second objective of this work is to develop a system which is able to forecast the *PR* of the PV module using minimal and understandable variables. The main purpose of this system is not only to know in advance what the approximate performance will be in the coming days, but also, if over several days the performance achieved is lower than predicted (and getting worse), it may be a sign of a breakdown or other problems in the installation.

This is to be done by using variables that are easily accessible to the average citizen. In particular, the performance depends closely on the type of day it is going to be, as well as the ambient temperature, values that are easily obtainable through weather forecast websites, or with simple observation and a thermometer if the day has passed. Specifically, the 12 classes of days used by the classifier described above (variable *DayClass*), and the maximum ambient temperature reached during the day (variable T_{max}) will be used.

For the database, out of the 9 years of measurements, the first seven years will be taken to train the system (from 1 January 2011 to 31 December 2017) and then the measurements of the following two years (from 1 January 2018 to 4 February 2020) will be used to check whether the system would have predicted the *PR* values correctly.

When developing the system, as was done with the classifier, a grid-type system is chosen, using the Wang-Mendel algorithm, and then tuned using the Levenberg-Marquardt algorithm, using the *xfdm* and *xfsl* tools of Xfuzzy, respectively. For the *DayClass* variable, 12 membership functions are chosen (which cannot be reduced because they are the 12 classes used by the classifier) and, for the T_{max} variable, from 10 membership functions onwards the RMSE is not significantly improved, obtaining inappropriate results if fewer functions are chosen. The resulting functions for T_{max}, after training, are shown in Fig. 10a.

The resulting matrix of the rule-base is 12x10. Although this results in 120 possible rules, in reality only 82 are generated, as there are cases within the matrix that never occur (such as, for example, that on a *Cloudy_L* type day, *Hot_VH* type temperatures are reached at T_{max}). The 82 singletons generated for the output of the system, PR_{ini}, can be simplified, without significantly worsening the behaviour of the system. Concretely, by clustering, these 82 singletons are grouped into 15, obtaining the final result shown in Fig. 10b. In this way, the rule-base illustrated in Fig. 11 is reached. Finally, this rule-base can be simplified, as was done with the classifier, by using the Fuzzy Tabular Simplification algorithm, arriving at a final reduced system of 15 rules. The application of the clustering simplification of the PR_{ini} singletons, as well as the tabular simplification of the rule-base, has been done by using the Xfuzzy tool *xfsp*.

This initial forecasting system (called *Prediction*), on its own, does not achieve satisfactory forecasting values. This is because there is another factor, also easily accessible by an average citizen, which affects the performance and that is the time of use of the PV module, since, as time goes by, the PV installation degrades and the performance decreases. Therefore, the PR_{ini} output of the *Prediction* module is connected, in cascade, with a new module (called *Correction*), in which a new input variable called t_{use} also enters, which is the time of use of the PV module in months, achieving the hierarchical structure shown in Fig. 12.

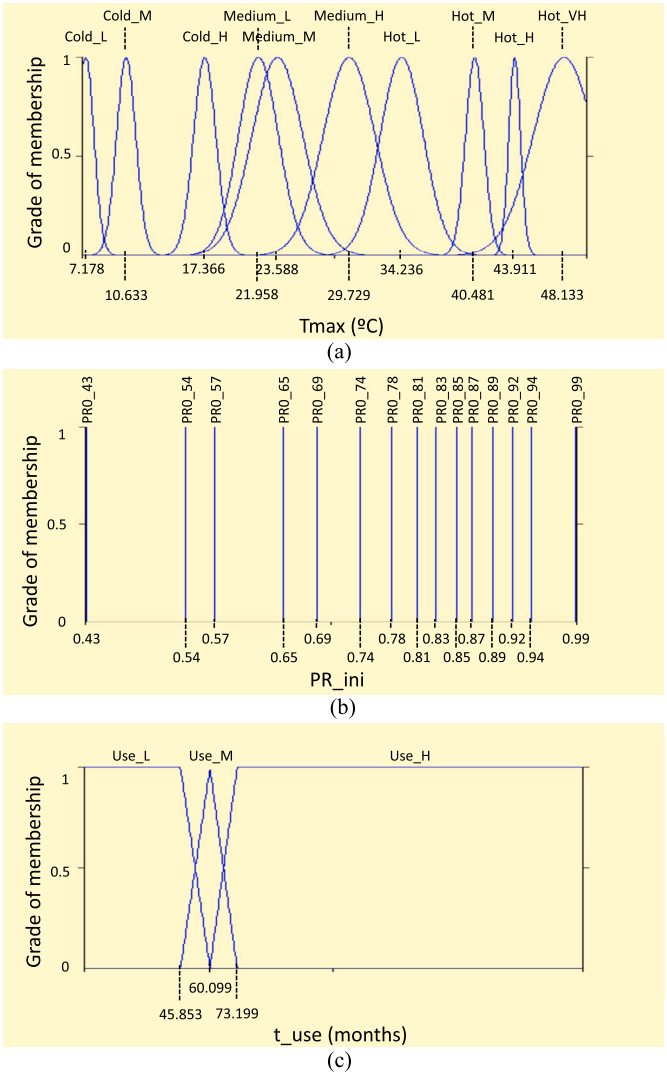

Fig. 10. (a) Membership functions for the T_{max} input; (b) Membership functions for the PR_{ini} output; (c) Membership functions for the t_{use} input.

For the *Correction* module, the system that generates the best answer is of the first-order Takagi-Sugeno-Kang (TSK) type [62][63]. The whole system, consisting of *Prediction* (previously trained) and *Correction*, is trained using the Levenberg-Marquardt algorithm. From 3 membership functions onwards for the variable t_{use}, the RMSE of the total system does not improve, reaching, after training, the final functions shown in Fig. 10c. Therefore, to the 15 rules of the *Prediction* module, 3 rules of the *Correction* module are added, resulting in a total system of 18 fairly reduced rules.

5.1. Results and discussion of the forecasting system

Before checking the good performance of the Neuro-Fuzzy Forecasting System (NFFS), proposed in this paper, the RMSE achieved by this will first be compared with that achieved with other training algorithms. For this comparison, the Linear Regression, Multi Layer Perceptron (MLP), Radial Basis Function (RBF) Network, M5P and M5Rules algorithms have been used. Linear regression (LR) works by estimating coefficients for a line or hyperplane that best fits the training data [64]; MLP is an artificial neural network (ANN) that uses backpropagation to learn a multi-layer perceptron to classify instances [65]; RBFNetwork implements a normalized Gaussian radial basis function network that

DayClass	Tmax									
	Cold_L	Cold_M	Cold_H	Medium_L	Medium_M	Medium_H	Hot_L	Hot_M	Hot_H	Hot_VH
Cloudy_L	PR0_54	PR0_57	PR0_57	PR0_57						
Cloudy_M		PR0_74	PR0_69	PR0_65						
Cloudy_H	PR0_43	PR0_78	PR0_81	PR0_81	PR0_78	PR0_83	PR0_69			
Sunny_L		PR0_81	PR0_81	PR0_81	PR0_78		PR0_89			
Sunny_M		PR0_89	PR0_87	PR0_89	PR0_83	PR0_85	PR0_83	PR0_83	PR0_85	PR0_83
Sunny_H			PR0_94	PR0_87	PR0_94	PR0_89	PR0_87	PR0_85	PR0_85	PR0_83
ParSunny_L		PR0_74	PR0_83	PR0_83	PR0_78	PR0_85	PR0_85	PR0_83		
ParSunny_M		PR0_89	PR0_87	PR0_85	PR0_89	PR0_87	PR0_85	PR0_83	PR0_83	PR0_83
ParSunny_H				PR0_89	PR0_94	PR0_87	PR0_87	PR0_87	PR0_83	
ParCloudy_L		PR0_83	PR0_83	PR0_81	PR0_83	PR0_83	PR0_81	PR0_81	PR0_78	
ParCloudy_M		PR0_85	PR0_87	PR0_89	PR0_85	PR0_85	PR0_85	PR0_87	PR0_83	
ParCloudy_H		PR0_99	PR0_92	PR0_89	PR0_89	PR0_85	PR0_87	PR0_85	PR0_85	

Fig. 11. *Prediction* module rule-base.

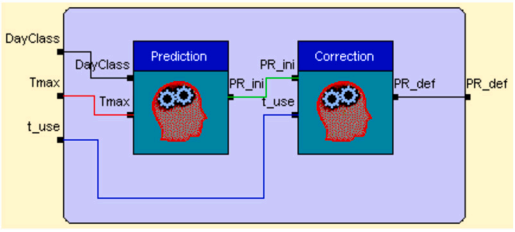

Fig. 12. Forecasting system structure.

Table 6
Comparison of performance assessment parameters for various approaches to the forecasting problem.

Approaches	Training data set		Testing data set	
	RMSE	MAE	RMSE	MAE
LR	4.304%	3.115%	6.055%	4.246%
MLP	4.678%	3.874%	6.059%	4.307%
RBFNetwork	5.190%	3.661%	8.130%	5.723%
M5P	3.188%	2.236%	6.917%	4.969%
M5Rules	3.360%	2.383%	6.718%	4.740%
NFFS	**3.021%**	**2.187%**	**5.961%**	**4.175%**

uses the k-means clustering algorithm to provide the basis functions and learns a linear regression on top of that. Moreover, symmetric multivariate Gaussians are fit to the data from each cluster [65]; M5P is an algorithm in trees category which implements the original M5 algorithm [66][67] for generating trees. In a similar way, M5Rules is the implementation of algorithm M5 for generating rules [68]. For the implementation of these algorithms, the WEKA software has been used, using the default parameters.

After applying all the algorithms mentioned above, the performance assessment parameters achieved, both for the training data of the first 7 years and for the test of the last 2 years, are shown in Table 6, where RMSE stands for Root-Mean-Square deviation, and MAE for Mean Absolute Error.

As can be seen in the table, of all the methods used, the one that provides the best training and test results is the one proposed in this work.

To visualise more clearly the behaviour achieved, Fig. 13 illustrates the *PR* values predicted by the system together with the actual *PR* values achieved during the 7 years of training. Two things can be observed in this figure: firstly, that the performance of the installation is indeed gradually decreasing over the years; secondly, that the system has acquired a quite satisfactory behaviour, including the degradation over time, mentioned above.

The next step to be checked is the behaviour of the system over the last two years, which are the data used as test data, shown in Fig. 14. In this case, it can be seen that the predicted behaviour fits reasonably well with the real one except in one area, which corresponds to the t_{use} values between 101 and 104 months, corresponding to the summer of

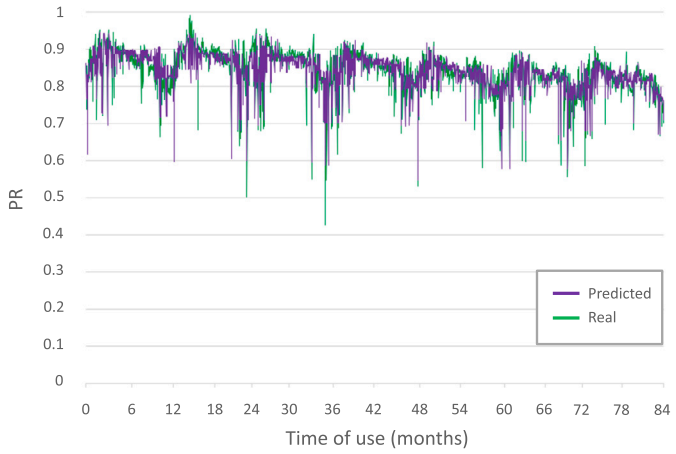

Fig. 13. *PR* values provided by the neuro-fuzzy forecasting system, versus actual *PR* values achieved with the training data (from 1 January 2011 to 31 December 2017).

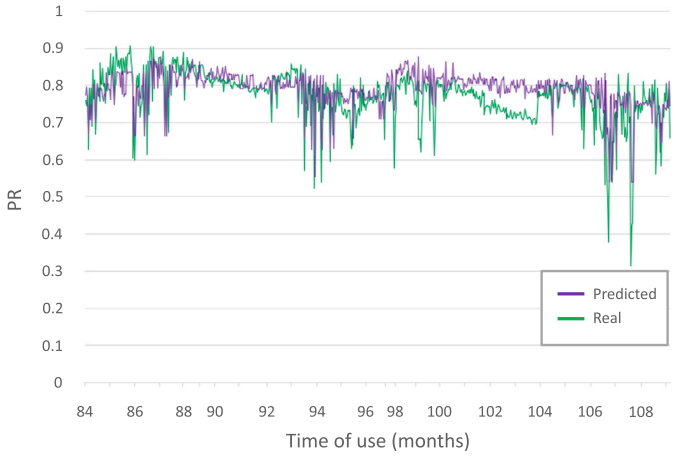

Fig. 14. *PR* values provided by the neuro-fuzzy forecasting system, versus actual *PR* values achieved with the test data (from 1 January 2018 to 4 February 2020).

2019. The reason for this difference is due to several motives. Firstly, there is more dust accumulation in summer than in other seasons, due to the almost total absence of rainfall, as well as the arrival of Saharan dust. Secondly, in the summer of 2019 there was an abnormal increase in air pollution in Cordoba [69], increasing the accumulation of dirt on the panels compared to other years. Finally, in previous years, the solar panels were cleaned more regularly during the summer, however, during the months of May, June, July, and August 2019 they were not cleaned at all. It was not until 1 September 2019 (just after the summer holidays) that the solar panels were cleaned. From that date onwards, the predicted and actual values converge again.

Therefore, the system not only serves to predict the performance values, but, as in the case described above, it also serves to detect anomalous behaviour in the system, so that those responsible for the maintenance of the installation could be alerted to these cases and react accordingly.

6. Conclusions

In this paper, two neuro-fuzzy systems have been presented: one is responsible for classifying daily irradiance profiles as a human would and the other is responsible for forecasting the performance of the PV installation according to the type of day expected, the maximum ambient temperature and the time of use of the installation.

In the case of the classifier, the classes provided are *Cloudy, Partly Cloudy, Partly Sunny* and *Sunny*, which, in turn, are divided into *Low, Medium* and *High* subclasses each one of them.

Once the parameters to be used have been determined and simplification techniques have been applied, such as the use of incomplete rule bases or Fuzzy Tabular Simplification, a final system of only 12 rules is obtained, which in turn is linguistically interpretable. Although the system obtained is extremely simple, it responds adequately, classifying the different daily irradiance profiles as expected.

With the classifier completed and tested for correct operation, a large number of days recorded in the south of Spain are classified, studying the performances obtained for each of the classes, as well as the temperatures recorded in them. From this study, it is determined that the classes of days with the best *PR* are those with a high irradiation value, with the classes with passing clouds prevailing over those with clear days. The latter is mainly due to the fact that there are less losses when the modules have a lower temperature and by the compensation of instantaneous irradiance values due to the *Cloud Enhancement* phenomenon.

With this large database of classified days, a neuro-fuzzy system has been designed that is capable of forecasting the *PR* of the PV system with variables that are easily predictable by an average citizen, adding, to the day class, the variables of maximum ambient temperature reached during the day, as well as the time of use of the PV installation. Specifically, a hierarchical system is used in which a first system predicts the *PR* with the day class and the maximum temperature and this first value is corrected with a second module in which the time of use of the PV modules is also included.

From this large database, the first 7 years of measurements have been used to train the system and the last 2 years to test it. With the proposed technique, a reduced system of only 18 rules is reached, which achieves quite satisfactory predictive values, not only emulating the gradual drop in performance due to the degradation of the time of use, but it has also served to detect anomalous behaviour in the installation during the summer of 2019.

Funding

This work was supported in part by the project Monitoring And Integration of energy data with Seamless Temporal Accuracy for photovoltaic plants (MISTA), PID2019-108953RA-C22, funded by the Spanish Ministry of Economy and Competitiveness.

CRediT authorship contribution statement

Andrés Gersnoviez: Conceptualization, Investigation, Methodology, Supervision, Writing – original draft. **Juan C. Gámez-Granados:** Formal analysis, Software, Writing – review & editing. **Marta Cabrera-Fernández:** Formal analysis, Software, Validation. **Isabel Santiago:** Data curation, Resources, Writing – review & editing. **Eduardo Cañete-Carmona:** Formal analysis, Resources, Writing – review & editing. **María Brox:** Investigation, Supervision, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors would like to thank the company Solar del Valle SL, who, thanks to their collaboration agreements with our research group, provided us with the data monitored in the PV installation; the authors would also like to thank the Down Syndrome Association in Córdoba, Spain, who are the owners of the PV installation under study, and with which the University of Córdoba granted the UCO Social Innova project called SEnseable Nursing homeS (SENS); finally, the authors would like to thank AEMET for providing the sky state data to be compared with the results of the classifier.

References

[1] W. Omran, Performance analysis of grid-connected photovoltaic systems, PhD Dissertation, 2010.
[2] B. Hartmann, Comparing various solar irradiance categorization methods – a critique on robustness, Renew. Energy 154 (2020) 661–671, https://doi.org/10.1016/j.renene.2020.03.055.
[3] D.H.W. Li, J.C. Lam, An analysis of climatic parameters and sky condition classification, Build. Environ. 36 (4) (2001) 435–445, https://doi.org/10.1016/S0360-1323(00)00027-5.
[4] M. Muselli, P. Poggi, G. Notton, A. Louche, Classification of typical meteorological days from global irradiation records and comparison between two Mediterranean coastal sites in Corsica Island, Energy Convers. Manag. 41 (10) (2000) 1043–1063, https://doi.org/10.1016/S0196-8904(99)00139-9.
[5] E. Aranovitch, C. Gandino, D. Gillaert, Elaboration of synthetic climatic data for the determination of the performance of solar systems, in: Colloque Météorologie et Energies Renouvelables, 1984, pp. 315–338.
[6] R. Rahim Baharuddin, R. Mulyadi, Classification of daylight and radiation data into three sky conditions by cloud ratio and sunshine duration, Energy Build. 36 (7) (2004) 660–666, https://doi.org/10.1016/j.enbuild.2004.01.012.
[7] Baharuddin, S.S.Y. Lau, R. Rahim, Daylight availability in Hong Kong: classification into three sky conditions, Archit. Sci. Rev. 53 (4) (2010) 396–407, https://doi.org/10.3763/asre.2009.0084.
[8] J.S. Stein, C.W. Hansen, M.J. Reno, The variability index: a new and novel metric for quantifying irradiance and PV output variability, in: World Renew. Energy Forum, 2012.
[9] C. Trueblood, S. Coley, T. Key, L. Rogers, A. Ellis, C. Hansen, E. Philpot, PV measures up for fleet duty: data from a Tennessee plant are used to illustrate metrics that characterize plant performance, IEEE Power Energy Mag. 11 (2) (2013) 33–44, https://doi.org/10.1109/MPE.2012.2234405.
[10] J. Huang, A. Troccoli, P. Coppin, An analytical comparison of four approaches to modelling the daily variability of solar irradiance using meteorological records, Renew. Energy 72 (2014) 195–202, https://doi.org/10.1016/j.renene.2014.07.015.
[11] A. Avila, P.R. Vizcaya, R. Diez, Daily irradiance test signal for photovoltaic systems by selection from long-term data, Renew. Energy 131 (2019) 755–762, https://doi.org/10.1016/j.renene.2018.07.071.
[12] A. Maafi, S. Harrouni, Preliminary results of the fractal classification of daily solar irradiances, Sol. Energy 75 (1) (2003) 53–61, https://doi.org/10.1016/S0038-092X(03)00192-0.
[13] S. Harrouni, A. Guessoum, A. Maafi, Classification of daily solar irradiation by fractional analysis of 10-min-means of solar irradiance, Theor. Appl. Climatol. 80 (1) (2005) 27–36, https://doi.org/10.1007/s00704-004-0085-0.
[14] S. Harrouni, Fractal classification of typical meteorological days from global solar irradiance: application to five sites of different climates, in: Modeling Solar Radiation at the Earth's Surface, 2008, pp. 29–54.
[15] A. Louche, G. Notton, P. Poggi, G. Simonnot, Classification of direct irradiation days in view of energetic applications, Sol. Energy 46 (4) (1991) 255–259, https://doi.org/10.1016/0038-092X(91)90071-4.
[16] M. Nijhuis, B.G. Rawn, M. Gibescu, Classification technique to quantify the significance of partly cloudy conditions for reserve requirements due to photovoltaic plants, in: 2011 IEEE Trondheim PowerTech, 2011, pp. 1–7.
[17] L. Fortuna, G. Nunnari, S. Nunnari, A new fine-grained classification strategy for solar daily radiation patterns, Pattern Recognit. Lett. 81 (2016) 110–117, https://doi.org/10.1016/j.patrec.2016.03.019.
[18] D. Djafer, A. Irbah, M. Zaiani, Identification of clear days from solar irradiance observations using a new method based on the wavelet transform, Renew. Energy 101 (2017) 347–355, https://doi.org/10.1016/j.renene.2016.08.038.

[19] M. Gastón-Romeo, T. Leon, F. Mallor, L. Ramírez-Santigosa, A morphological clustering method for daily solar radiation curves, Sol. Energy 85 (9) (2011) 1824–1836, https://doi.org/10.1016/j.solener.2011.04.023.
[20] G. Chicco, V. Cocina, F. Spertino, Characterization of solar irradiance profiles for photovoltaic system studies through data rescaling in time and amplitude, in: Proc. 49th Int. Univ. Power Eng. Conf. (UPEC'2014), September, 2014, pp. 1–6.
[21] A.A. Munshi, Y.A.-R.I. Mohamed, Photovoltaic power pattern clustering based on conventional and swarm clustering methods, Sol. Energy 124 (2016) 39–56, https://doi.org/10.1016/j.solener.2015.11.010.
[22] T. Soubdhan, R. Emilion, R. Calif, Classification of daily solar radiation distributions using a mixture of Dirichlet distributions, Sol. Energy 83 (7) (2009) 1056–1063, https://doi.org/10.1016/j.solener.2009.01.010.
[23] S. Alimohammadi, D. He, Multi-stage algorithm for uncertainty analysis of solar power forecasting, in: Proc. 2016 IEEE Power Energy Soc. Gen. Mtg. (PESGM'2016), July, 2016, pp. 1–5.
[24] J. Calbó, J.A. González, D. Pagè, A method for sky-condition classification from ground-based solar radiation measurements, J. Appl. Meteorol. 40 (12) (2001) 2193–2199, https://doi.org/10.1175/1520-0450(2001)040<2193:AMFSCC{&gt2.0.CO;2}.
[25] N. Umemiya, T. Kanou, Classification of sky conditions by the ranges of insolation indices considering CIE standard for general sky, J. Light Vis. Environ. 32 (1) (2008) 14–19, https://doi.org/10.2150/jlve.32.14.
[26] S. Moreno-Tejera, M.A. Silva-Pérez, L. Ramírez-Santigosa, I. Lillo-Bravo, Classification of days according to DNI profiles using clustering techniques, Sol. Energy 146 (2017) 319–333, https://doi.org/10.1016/j.solener.2017.02.031.
[27] B.O. Kang, K.-S. Tam, A new characterization and classification method for daily sky conditions based on ground-based solar irradiance measurement data, Sol. Energy 94 (2013) 102–118, https://doi.org/10.1016/j.solener.2013.04.007.
[28] C. Sacré, Automatic classification of meteorological days for energetic applications, in: Colloque Météorologie et Energies Renouvelables, 1984, pp. 615–630.
[29] F. Fabero, M. Alonso-Abella, F. Chenlo, Influence of irradiation variations on PV systems at differents time scales, in: Proc. 14th Eur. PV Sol. Energy Conf. Exhib., 1997, pp. 2299–2302.
[30] P. Boullier, M. Le Chapellier, Weather types: an improved analysis, in: Colloque Météorologie et Energies Renouvelables, 1984, pp. 632–653.
[31] S. Wittkopf, S. Valliappan, L. Liu, K.S. Ang, S.C.J. Cheng, Analytical performance monitoring of a 142.5kWp grid-connected rooftop BIPV system in Singapore, Renew. Energy 47 (2012) 9–20, https://doi.org/10.1016/j.renene.2012.03.034.
[32] I. Santiago, J.L. Esquivel-Martin, D. Trillo-Montero, R.J. Real-Calvo, V. Pallarés-López, Classification of irradiance daily profiles and the behavior of a photovoltaic plant elements: the effects of cloud enhancement, Appl. Sci. 11 (11) (2021) 5230, https://doi.org/10.3390/app11115230.
[33] L.A. Zadeh, Outline of a new approach to the analysis of complex systems and decision processes, IEEE Trans. Syst. Man Cybern. SMC-3 (1) (1973) 28–44, https://doi.org/10.1109/TSMC.1973.5408575.
[34] L. Suganthi, S. Iniyan, A.A. Samuel, Applications of fuzzy logic in renewable energy systems – a review, Renew. Sustain. Energy Rev. 48 (2015) 585–607, https://doi.org/10.1016/j.rser.2015.04.037.
[35] J.-S.R. Jang, C.-T. Sun, E. Mizutani, Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence, Prentice Hall, Upper Saddle River, NJ, USA, 1997.
[36] M. Yesilbudak, M. Çolak, R. Bayindir, A review of data mining and solar power prediction, in: Proc. 2016 IEEE Int. Conf. Renew. Energy Res. Appl. (ICRERA'2016), 2016, pp. 1117–1121.
[37] D. Patel, S. Patel, P. Patel, M. Shah, Solar radiation and solar energy estimation using ANN and fuzzy logic concept: a comprehensive and systematic study, Environ. Sci. Pollut. Res. 29 (2022) 32428–32442, https://doi.org/10.1007/s11356-022-19185-z.
[38] K. Kumar Ranjith, M. Surya Kalavathi, Artificial intelligence based forecast models for predicting solar power generation, Mater. Today Proc. 5 (1) (2018) 796–802, https://doi.org/10.1016/j.matpr.2017.11.149.
[39] Z.P. Ncane, A.K. Saha, Forecasting solar power generation using fuzzy logic and artificial neural network, in: Proc. 2019 South. African Univ. Power Eng. Conf. Robot. Mechatron./Pattern Recognit. Assoc. South Africa (SAUPEC/RobMech/PRASA'2019), 2019, pp. 518–523.
[40] O. Yazdanbasksh, A. Krahn, S. Dick, Predicting solar power output using complex fuzzy logic, in: Proc. 2013 IFSA World Congr. NAFIPS Annu. Meet. (IFSA/NAFIPS'2013), 2013, pp. 1243–1248.
[41] J. Tavoosi, A.A. Suratgar, M.B. Menhaj, A. Mosavi, A. Mohammadzadeh, E. Ranjbar, Modeling renewable energy systems by a self-evolving nonlinear consequent part recurrent type-2 fuzzy system for power prediction, Sustainability 13 (6) (2021) 3301, https://doi.org/10.3390/su13063301.
[42] M. Rizwan, M. Jamil, S. Kirmani, D.P. Kothari, Fuzzy logic based modeling and estimation of global solar energy using meteorological parameters, Energy 70 (2014) 685–691, https://doi.org/10.1016/j.energy.2014.04.057.
[43] S. Jafarzadeh, M.S. Fadali, C.Y. Evrenosoglu, Solar power prediction using interval type-2 TSK modeling, IEEE Trans. Sustain. Energy 4 (2) (2013) 333–339, https://doi.org/10.1109/TSTE.2012.2224893.
[44] D. Trillo-Montero, I. Santiago, J.J. Luna-Rodríguez, R. Real-Calvo, Development of a software application to evaluate the performance and energy losses of grid-connected photovoltaic systems, Energy Convers. Manag. 81 (2014) 144–159, https://doi.org/10.1016/j.enconman.2014.02.026.
[45] M.J. Gacto, R. Alcalá, F. Herrera, Interpretability of linguistic fuzzy rule-based systems: an overview of interpretability measures, Inf. Sci. 181 (20) (2011) 4340–4360, https://doi.org/10.1016/j.ins.2011.02.021.
[46] I. Baturone, A. Gersnoviez, Automatic extraction of linguistic models for image description, in: Proc. IEEE 2010 Int. Conf. Fuzzy Syst. (FUZZ-IEEE'2010), 2010, pp. 1–8.
[47] A. Gersnoviez, I. Baturone, Rule simplification method based on covering indexes for fuzzy classifiers, in: Proc. IEEE 2021 Int. Conf. Fuzzy Syst. (FUZZ-IEEE'2021), 2021, pp. 1–6.
[48] H. Ishibuchi, T. Nakashima, Effect of rule weights in fuzzy rule-based classification systems, IEEE Trans. Fuzzy Syst. 9 (4) (2001) 506–515, https://doi.org/10.1109/91.940964.
[49] L.-X. Wang, J.M. Mendel, Fuzzy basis functions, universal approximation, and orthogonal least-squares learning, IEEE Trans. Neural Netw. 3 (5) (1992) 807–814, https://doi.org/10.1109/72.159070.
[50] R. Battiti, First- and second-order methods for learning: between steepest descent and Newton's method, Neural Comput. 4 (2) (1992) 141–166, https://doi.org/10.1162/neco.1992.4.2.141.
[51] Xfuzzy: fuzzy logic design tools, available at http://www.imse.cnm.es/Xfuzzy.
[52] I. Baturone, F.J. Moreno-Velo, A. Gersnoviez, A CAD approach to simplify fuzzy system descriptions, in: Proc. IEEE 2006 Int. Conf. Fuzzy Syst. (FUZZ-IEEE'2006), 2006, pp. 2392–2399.
[53] N. Chinchor, MUC-4 evaluation metrics, in: Proc. Fourth Message Understanding Conference (MUC-4), 1992, pp. 22–29.
[54] J.-S.R. Jang, ANFIS: adaptive-network-based fuzzy inference system, IEEE Trans. Syst. Man Cybern. 23 (3) (1993) 665–685, https://doi.org/10.1109/21.256541.
[55] J.R. Quinlan, C4.5: Programs for Machine Learning, Elsevier, 2014.
[56] E. Frank, M.A. Hall, I.H. Witten, The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques, 4th ed., Morgan Kaufmann, Burlington, 2016.
[57] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, ACM SIGKDD Explor. Newsl. 11 (1) (2009) 10–18, https://doi.org/10.1145/1656274.1656278.
[58] I. Santiago, D. Trillo-Montero, I.M. Moreno-García, V. Pallarés-López, J.J. Luna-Rodríguez, Modeling of photovoltaic cell temperature losses: a review and a practice case in South Spain, Renew. Sustain. Energy Rev. 90 (2018) 70–89, https://doi.org/10.1016/j.rser.2018.03.054.
[59] R. Tapakis, A.G. Charalambides, Enhanced values of global irradiance due to the presence of clouds in Eastern Mediterranean, Renew. Energy 62 (2014) 459–467, https://doi.org/10.1016/j.renene.2013.08.001.
[60] G.H. Yordanov, A study of extreme overirradiance events for solar energy applications using NASA's I3RC Monte Carlo radiative transfer model, Sol. Energy 112 (2015) 954–965, https://doi.org/10.1016/j.solener.2015.10.014.
[61] M. Järvelä, K. Lappalainen, S. Valkealahti, Characteristics of the cloud enhancement phenomenon and PV power plants, Sol. Energy 196 (2020) 137–145, https://doi.org/10.1016/j.solener.2019.11.090.
[62] T. Takagi, M. Sugeno, Fuzzy identification of systems and its application to modeling and control, IEEE Trans. Syst. Man Cybern. SMC-15 (1985) 116–132, https://doi.org/10.1109/TSMC.1985.6313399.
[63] M. Sugeno, G.T. Kang, Structure identification of fuzzy model, Fuzzy Sets Syst. 28 (1) (1988) 15–33, https://doi.org/10.1016/0165-0114(88)90113-3.
[64] N.R. Draper, H. Smith, Applied Regression Analysis, John Wiley & Sons, New York, USA, 1981.
[65] S. Haykin, Neural Networks, a Comprehensive Foundation, 2nd ed., Prentice Hall, New Jersey, 1999.
[66] J.R. Quinlan, Learning with continuous classes, in: Proc. 5th Australian Joint Conf. Artif. Intell. (AJCAI'1992), 1992, pp. 343–348.
[67] Y. Wang, I.H. Witten, Induction of model trees for predicting continuous classes, in: Proc. 9th European Conference on Machine Learning (ECML'1997), 1997.
[68] G. Holmes, M. Hall, E. Prank, Generating Rule Sets from Model Trees, Lect Notes Comput. Sci., vol. 1747, 1999.
[69] M. in't Veld, C. Carnerero, J. Massagué, A. Alastuey, J.D. de la Rosa, A.M. Sánchez de la Campa, M. Escudero, E. Mantilla, G. Gangoiti, C. Pérez García-Pando, M. Olid, J.R. Moreta, J.L. Hernández, J. Santamaría, M. Millán, X. Querol, Understanding the local and remote source contributions to ambient O_3 during a pollution episode using a combination of experimental approaches in the Guadalquivir valley, southern Spain, Sci. Total Environ. 777 (2021), https://doi.org/10.1016/j.scitotenv.2020.144579.