Expert Systems With Applications xxx (xxxx) xxx

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Binding data mining and expert knowledge for one-day-ahead prediction of hourly global solar radiation

José del Campo-Ávila [a,*], Abdelatif Takilalte [b], Albert Bifet [c,d], Llanos Mora-López [a]

[a] Universidad de Málaga, Departamento de Lenguajes y Ciencias de la Computación, Campus de Teatinos, 29071 Málaga, Spain
[b] Centre de Développement des Energies Renouvelables, CDER, BP 62 Route de l'Observatoire, Bouzaréah, 16340, Algiers, Algeria
[c] LTCI, Télécom Paris, 46 rue Barrault, Paris, Cedex 13 75634, France
[d] University of Waikato, Hamilton, New Zealand

ARTICLE INFO

Keywords:
Data mining
One-day-ahead prediction
Hourly global solar radiation
Expert systems

ABSTRACT

A new methodology to predict one-day-ahead hourly solar global radiation is proposed in this paper. This information is very useful to address many real problems; for instance, energy-market decision making is one of the contexts where that information is essential to ensure the correct integration of grid-connected photovoltaic solar systems. The developed methodology is based on the contribution of different experts to obtain improved data-driven models when included in the data mining process. The modelling phase, when models are induced and new patterns can be identified, is the one that most benefits from that expert knowledge. In this case, it is achieved by combining clustering, regression and classification methods that exploit meteorological data (directly measured or predicted by weather services). The developed models have been embedded in a prediction system that offers reliable forecasts on next-day hourly global solar radiation. As a result of the automatic learning process including the knowledge of different experts, 14 different types of day were identified based on the shape of hourly solar radiation throughout a day. The conventional definitions of types of days, that usually consider 4 options, are updated with this new proposal. The next-day prediction of hourly global radiation is obtained in two phases: in the first one, the next-day type is obtained from among the 14 possible types of day; in the second one, values of hourly global radiation are obtained using the centroid of the predicted type of day and extraterrestrial solar radiation. The relative root mean square error of the prediction model is less than 20%, meaning a significant reduction compared to previous models. Moreover, the proposed models can be recognized in the context of eXplainable Artificial Intelligence.

1. Introduction

Data mining techniques have proven to be very useful in the search for solutions to complex problems involving many variables and large datasets. These types of problems are difficult to tackle with traditional data analysis techniques, and, for this reason, there is currently a great demand for intelligent systems.

Nowadays, data sources are more frequent, more comprehensive and more accessible. This is both an opportunity and a challenge for those seeking to extract knowledge from different domains. It is an opportunity as such an amount of information has never before been available, but it is also a challenge given the heterogeneity and huge amount of data.

Data are the core of any data mining process, and business understanding is present in different methodologies defined to guide such mining processes (Ponsard et al., 2017). Specifically, business (and data) understanding are in the early stages of the procedure. This reveals the importance of incorporating the knowledge of experts right from the start in order to discover new useful knowledge and to deploy high quality products. Experts provide their proficiency to start the data mining process and they can confirm the correctness of discovered patterns at an intermediate point, when models are generated (if they are explainable). Furthermore, they can discover new useful knowledge that would be very difficult, or impossible, to obtain given the great amount of data being handled. Finally, all that knowledge, if positively evaluated, could be used in real systems (to predict, to simulate, to assess in decision making, etc.).

The use of data mining techniques is therefore extending to different areas of application. Moreover, hybrid models have appeared in recent years, which combine the use of different techniques for the same problem, and which require the close collaboration of researchers from

* Corresponding author.
E-mail addresses: jcampo@uma.es (J. del Campo-Ávila), a.takilalte@cder.dz (A. Takilalte), abifet@waikato.ac.nz (A. Bifet), llanos@uma.es (L. Mora-López).

https://doi.org/10.1016/j.eswa.2020.114147
Received 24 June 2020; Received in revised form 17 October 2020; Accepted 20 October 2020
Available online 27 October 2020
0957-4174/

different fields for their correct development and implementation. The integration of the knowledge of different experts in phases of the data mining process can result in improved results compared to those obtained automatically.

1.1. Importance of predicting solar radiation

One of the areas in which these techniques are being successfully used is in the prediction of different meteorological parameters. Among these, the short-term prediction of solar global radiation is a key issue being approached with these models. This type of data is required to predict and evaluate the performance of solar energy system using solar radiation as the resource. Specifically, the prediction horizon of 24 h ahead is required for decision making on the energy market in order to integrate grid-connected large photovoltaic (PV) solar systems, as PV power estimation is relevant for distribution System Operators, energy traders and aggregators (Pierro et al., 2017). It is also important for PV self-consumption facilities, as knowing their production in advance can help to achieve an optimization of self-consumption (energy that is generated by the PV system and directly consumed in the house) (Ayala-Gilardón et al., 2018). In both cases, this prediction can help to improve their profitability and integration in the power grid. The production of such facilities depends on the configuration of the PV system (peak power, orientation and technology) and on the temperature and solar radiation received. Therefore, knowing in advance the solar radiation received is fundamental to be able to predict their production.

1.2. Previous approaches to predict solar radiation

Different approaches have been proposed for predicting solar global radiation: data mining models, cloud motion tracking, numerical weather predictions and hybrid models. The input variables required and the complexity of these approaches are also very different. Moreover, it is also important to consider the forecast horizon for which these models have been developed. Most of the models for predicting hourly solar global radiation predict for next hour, that is, the forecast horizon is equal to one hour (Blaga et al., 2019). In these cases, errors are lower than those errors achieved for longer forecast horizons such as for next-day.

As regards previous research into the use of data mining models to predict solar radiation, special mention can be made of the following using different techniques: artificial neural networks or hybrid models proposed in Gairaa et al. (2016), Jiménez-Pérez and Mora-López (2016), Krakovsky and Luzgin (2018) and Ozgoren et al. (2012); support vector machines proposed in Bektas (2014) and Deo et al. (2016); and fuzzy models, as in Kisi (2014). Some of these studies face the problem of predicting for one-step-ahead (next hour) while others are focused on predicting values in advance, where the forecast horizon can range from several hours to one or more days.

As expected, prediction errors are greater as the forecast time horizon increases (Blaga et al., 2019). While the mean rRMSE is 23.8% for one-hour horizon, the mean rRMSE is 42.22% for one-day ahead horizon for all the models reported in Blaga et al. (2019). If only data mining models are considered, the mean values of rRMSE are approximately 22% and 40% for one-hour and one-day horizon, respectively.

Regarding the methods that predict hourly values of solar global radiation for next day (one-day horizon), this type of prediction is little discussed in the literature, as is pointed out in Blaga et al. (2019). Among the models that induce ANNs, in Mellit and Pavan (2010) an MLP (multilayer perceptron) using values of the mean daily solar irradiance and air temperature as input is proposed; the model is trained with data from Trieste (Italy) and the rRMSE reported is 67%. ANN models are also used in Marquez and Coimbra (2011); up to eleven predicted meteorological variables, obtained from the US National Weather Services forecasting database, are used as input, together with two geometric/temporal variables; rRMSE values for one-day ahead predictions range between 20% and 23% (these data were obtained from figure 4.a in that paper, since the exact values are not in the text). In Voyant et al. (2013), authors report a rRMSE for MLP (multilayer perceptron) equal to 27.8% using endogenous variables and 27.3% when exogenous variables are also used (hourly pressure and cloudiness of the last day and daily average nebulosity of the two previous days); they used data from 5 French cities.

In a similar way, there are several works that predict energy generated by PV systems. For instance, several data mining models are evaluated to predict power forecasting in Rana and Rahman (2020). The forecast skill they obtain range between 0.23 and 0.30; values are estimated from 5 min to 3 h ahead. In Chen et al. (2011), the rMAE (%) obtained when predicting PV power range from 6.31 to 37.23 depending on the type of day (sunny, cloudy and rainy).

The idea of using clearness index for solar radiation clustering and to use the obtained clusters as a method to obtain different type of days has been previously used in Jiménez-Pérez and Mora-López (2016) and Monjoly et al. (2019). In both papers, only 4 different types of day were used; these 4 types of day were estimated without using expert knowledge. On the other hand, Muselli et al. (2000) proposes using 3 typical meteorological days classes, which is valid for representing the long-term performance of sky. Nevertheless, as it is explained later in this paper, the variability of observations in each cluster is so high that it results in great differences in the estimated values of hourly global radiation.

Some shortcomings can be observed in the previous approaches. For example, the models accuracy could be improved, the knowledge of different types of day is insufficient, the range of forecasting usually does not cover one-day-ahead and the applicability to different scenarios is limited.

1.3. Contributions and organization of the paper

In this paper, we present the advances achieved by combining the expert knowledge inside a data mining process with the objective of obtaining one-day-ahead prediction of hourly global solar radiation. Contributions can be perceived from different perspectives, but the most relevant are the following:

- Description of 14 different types of day, according to the shape of the hourly clearness index values. They have been identified using clustering methods and considering knowledge provided by experts. Therefore, the conventional definitions of types of days, that usually consider 3 or 4 options, are updated with this new proposal. Experimental results and expert advice suggest that it is a valid approximation.
- A methodology that can learn the type of day for the next day using meteorological and atmospheric data from the present day. By knowing the type of day, the one-day-ahead hourly global solar radiation can be predicted. This methodology has been tested with data from 10 different locations collected during 11 years, so it can be used in general case, but it could be adjusted with new data if the context differs considerably (or changes in the future).
- New data-driven models to predict one-day-ahead hourly solar radiation. These models have a double value:
  - Classification model itself reveals relations that are interesting for experts. It is represented with one model from which it is possible to extract understandable patterns. Thus, experts detect the importance of daily clearness index, humidity or cloudiness and how they are related.
  - Embedding these models in a prediction system can offer reliable forecasts for next-day hourly global solar radiation. This is very useful as described in Section 1.1.

The rest of the paper is organized as follows. Background knowledge used in this work is described in Section 2. It includes both the description and expressions to estimate one of the variables used (clearness index), the description of clustering and classification techniques used and the metrics and statistical validation proposed for evaluating the model. The proposed methodology is explained in Section 3. The experimental design is described in Section 4, including the description of the dataset, the software and the different algorithms used. Section 5 presents the results of each phase and the results for the whole proposed methodology. It also includes a comparison with the results obtained in previous works. Finally, Section 6 summarizes the main conclusions of the work.

2. Background and preliminaries

This section provides an overview of the background knowledge used in the proposal. First, we describe the atmospheric parameters characterizing hourly solar radiation. We then describe different methods for unsupervised and supervised learning, that are fundamental in the data mining process that we have conducted. Additionally, some measures and statistics to validate the quality of the models are enumerated.

2.1. Atmospheric parameters

We propose to use the hourly clearness index (k_h) in the analysed models as it allows us to remove the seasonal and daily trends observed in hourly solar global radiation. This variable is estimated using the hourly global solar radiation (G_h) and the extraterrestrial hourly solar radiation ($G_{0,h}$) according to the following expression:

$$k_h = \frac{G_h}{G_{0,h}}$$

$G_{0,h}$ is obtained using the expression:

$$G_{0,h} = I_{sc} E_0 \sin\alpha = I_{sc} E_0 (\sin\delta \sin\phi + \cos\delta \cos\phi \cos\omega_h)(W\,hm^{-2}),$$

where I_{sc} is the solar constant, E_0 is the eccentricity factor, α is the solar elevation, δ is the declination angle, ϕ is the latitude, ω_s is the hour angle centred at $(\omega_h - \pi/24, \omega_h + \pi/24)$. The definition of $G_{0,h}$ and the expressions to estimate E_0, δ and ω_h can be found in Iqbal (1983).

Considering that every day registers consecutive values of hourly radiation (namely, hourly clearness index), a set of curves can be built (one for every day and location). According to the shape and height of those curves, it is possible to compute new characteristics that allows the different types of day to be compared and fixed . That can be done by calculating the Area Under the Curve (AUC (2019)) computed by the trapezoidal rule as follows:

$$AUC = \frac{1}{2} \sum_{i=1}^{n-1} (x_{i+1} - x_i)(y_{i+1} + y_i - 2B)$$

where x_i is the value on horizontal axis, y_i is the value on vertical axis, n is the number of elements and B is the baseline value on the vertical axis.

2.2. Clustering methods

Clustering is one of the most important solutions when we need to discover relations in datasets that do not include knowledge about the classes. This type of learning is known as unsupervised learning. The goal is to form groups of examples sharing similar characteristics, but not with other groups.

A key aspect in this kind of learning is the definition of similarity functions (when working with qualitative attributes) or distance functions (when working with quantitative attributes). The most conventional distance functions are the Euclidean distance or the Pearson correlation distance in our domain, which consider numerical attributes (Xu & Tian, 2015).

There is a wide variety of conventional proposals that are based on different approaches (Xu & Tian, 2015): partitional (k-means, PAM, CLARA), hierarchical (BIRCH, CURE), fuzzy theory (FCM,FCS), density (DBSCAN), etc. One of the most important parameters to be configured is the number of clusters (or groups) to be considered in the process. Some may obviate this parameter but it is not a trivial decision.

In case of complex scenarios, like time-series, shape-based approaches consider the information in the whole time-series as individual objects and new distance functions (such as DTW Sakoe & Chiba, 1978) try to determine how different (or similar) they are Aghabozorgi et al. (2015).

Other clustering algorithms that can work with the shape of time-series are those based in kernels (like TGA kernels Cuturi, 2011). They make it easier to find groups in a high dimensional feature space and can work with arbitrary shapes, or separate overlapping clusters (Xu & Tian, 2015) .

2.3. Classification methods

Classification is an essential process in supervised learning. Its goal is to discover relations (like a mapping function) between the input data and the output data. The input data can be nominal (with discrete values) or numerical (with continuous values) while the output data are defined as a finite set of discrete values. That special output is called class (Liu & Wu, 2012). Regression methods are used when the class attribute is a numerical rather than nominal attribute.

There are multiple families of models and algorithms to induce those models. Some of them can even work with nominal and numerical classes. This subsection seeks to describe very well-known algorithms for most relevant types of models, focusing on classification.

However, it will be useful to describe two baseline methods before moving on to describe those algorithms. Selecting the *most frequent class* (or mode) in the dataset as predicted output is a straightforward way of ensuring a lower bound for the quality that should be surpassed by more advanced models. In the context of forecasting, the *persistence method* is another baseline (Perez-Ortiz et al., 2018). It supposes the prediction for a specific example is the class observed for the previous example.

The Naïve Bayes algorithm estimates the posterior probability of each class given an example ($P(class|example)$) from the dataset. When predicting, it selects the most likely class (Domingos & Pazzani, 1997). This algorithm performs very well even when the attributes are not conditionally independent.

Decision trees constitute a very common mode of representing the induced knowledge. The variety of methods to induce them is great and there are alternatives that can consider many different problems (nominal and numerical attributes, missing values, pre-pruned and post-pruned, etc.). One of the main characteristics of decision trees is their understandably. Every branch in the tree corresponds to a rule, and they are expressed in form of conditional statements that can be interpreted even by non-experts. C4.5 (Quinlan & Ross, 1993) is one of the best-known algorithms. It uses information gain to select the most convenient attribute for every decision to be taken in the expansion of the tree.

Classifying via regression is one alternative that allows the usage of methods in supervised learning that were not initially proposed for discrete classes, but rather numerical classes. There is an easy way to use regression with discrete classes: creating as many numeric indicators (new classes) as values defining the original class. Then, separate regression learners are trained to model the membership for every new numeric indicator. Model trees are thus proposed (Frank et al., 1998). They include regression functions at the leaves of the decision tree to solve a classification problem. If linear regression functions are replaced by logistic regression functions, we are talking about another method called logistic model trees (LMT) (Landwehr et al., 2005).

Artificial Neural Networks (ANNs) constitute another alternative to deal with classification problems. A neural network is defined by connecting different elements (called neurons or nodes) in a concrete way (input, internal layers and outputs), and then modifying the connection between those neurons (called weights) in order to model the relation between inputs and outputs. Neurons use activation functions that rule how they must behave depending on the input. A multilayer perceptron (MLP) is a feedforward artificial neural network (Popescu et al., 2009) that uses non-linear activation functions. It is characterized by several layers of nodes connected as a directed graph between the input and output layers.

Support Vector Machines (SVMs) form another group of methods to be used for classification. They use linear classifiers to determine the hyperplane that separates data in different categories, taking advantage of a previous transformation performed by generating a higher dimensional space. There are algorithms, such as SMO (Platt, 1998), that divide large problems in a sequence of smaller ones that are solved analytically, while allowing that SMO can be used to learn from larger training sets.

Multiple classifier systems, which can combine isolated base classifiers such as those mentioned above, benefit from the idea of using an ensemble of models to perform that classification task. They achieve highly accurate results, they are robust to noise and outliers and they do not overfit However, the results are more difficult to explain, because they are a combination of single models. Random forest is one ensemble method that achieves remarkable results (Wyner et al., 2017). It induces decision trees and uses a subset of attributes in every "new" model . Random forest, when using regression trees, can work on regression scenarios.

2.4. Error metrics for estimating hourly global solar radiation

The metric used for evaluating accuracy in classification models was estimated as the ratio of the number of misclassified instances, n, and the total number of instances, m according to the following expression:

$$Accuracy = \frac{n}{m} \cdot 100\ (\%)$$

The metrics used for evaluating the performance of the proposed methodology in the estimation of hourly solar global radiation had been previously described in Shcherbakov et al. (2013).

Let m be a set of values observed for the variable X. Let $X_t\,(t = 1, \ldots, m)$ represent each of those real values, $\bar{X}$ the mean of these values, and let $\hat{X}_t\,(t = 1, \ldots, m)$, represent the corresponding estimations. The expressions used are as follows:

- The mean absolute error (MAE) and the relative mean absolute error, ($rMAE$), are estimated using the expressions:

$$MAE = \frac{\sum_{t=1}^{m} |X_t - \hat{X}_t|}{m} \quad \text{and} \quad rMAE = \frac{MAE}{\bar{X}} \cdot 100\ (\%)$$

- The root mean square error, $RMSE$, and the relative (or normalized) root mean square error, $rRMSE$ estimated from the expressions:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{m} (X_t - \hat{X}_t)^2}{m}} \quad \text{and} \quad rRMSE = \frac{RMSE}{\bar{X}} \cdot 100\ \ (\%)$$

- The forecast skill over 24 h persistence forecasts, s, estimated as follows (Coimbra et al., 2013):

$$s = \left(1 - \frac{RMSE_{model}}{RMSE_{persistencemodel}}\right) \cdot 100\ \ (\%)$$

2.5. Statistical validation

In the search for the most suitable model to discover relations between input data and output data (class), it is important to estimate the average error that is produced by every considered option. Two tools – cross-validation and non-parametric statistical tests – can be used to obtain that estimation with statistical confidence.

Cross-validation is one of the most widely used methods to estimate prediction errors (Hastie et al., 2009), while avoiding optimistic overestimations. This method divides the dataset in k folds (or bins) and repeats k times a training-test process with k different training and test datasets. Every execution of the process uses one of the k folds of the original dataset as the test datasets, while the corresponding training datasets is formed with the rest of $k - 1$ folds. Thus, k executions are conducted and one average value can be estimated.

Statistical tests are used to statistically verify significant differences in results, in order to avoid spurious results and confirm that results are really different. There are parametric and non-parametric tests. The first tests assume that data follow a specific distribution while second ones do not consider that assumption. A non-parametric test, the Wilcoxon signed-ranks test, also known as paired samples Wilcoxon test, is proposed because of its simplicity and safeness when comparing two classifiers (Demšar, 2006).

3. Proposed methodology

This section presents the methodology used to predict hourly global solar radiation one-day ahead. This methodology is based on two hypotheses:

- It is possible to establish a limited number of type of days taking into account the shape of hourly clearness index.
- The type of next day can be estimated using meteorological parameters for present day and predictions of meteorological parameters for next day.

According to these two hypotheses, the methodology consists of two phases.

- The first phase determines the different types of days (number and shapes), according to the hourly clearness index (k_h). This task involves an unsupervised learning process, because there is no prior information about the classes (types of days). Data used in this phase are exclusively related to atmospheric parameters (radiation received before entering in the atmosphere $G_{0,h}$, radiation received in Earth's surface G_h and its ratio k_h).
- The second phase induces models that learn the type of day (shape) for next day, according to the meteorological and radiation information available for the present day. This is a supervised learning task because the types of days are known after the first phase and every example is labelled with its corresponding type of day.

Fig. 1 represents the division of methodology in two different phases and how they are combined to induce the models that capture the knowledge about next day hourly global solar radiation.

3.1. First phase (clustering)

Different pre-processing transformations and clustering methods have been tested in this phase. However, what is really important is the relevance of the knowledge of different experts that has steered the selection, as a wide automatically exploration does not conduct to results of sufficient quality, as will be discussed in Section 5. Fig. 2 is a schematic diagram of the process in this phase.

The only information used to estimate the types of days are the values of hourly clearness index (k_h). Radiation values such as ($G_{0,h}$ and

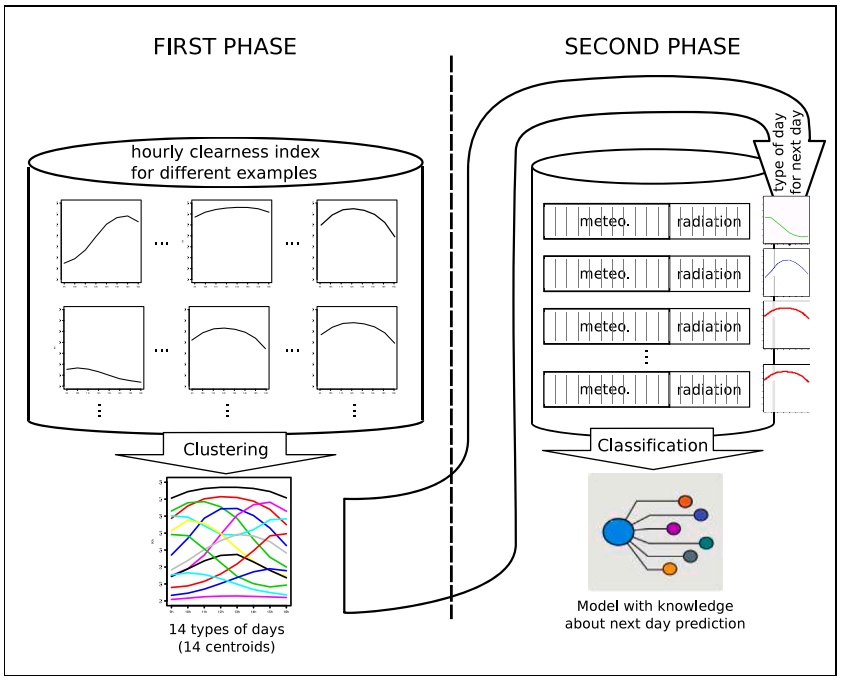

Fig. 1. Two phases defined to induce models that capture the knowledge about next day hourly global solar radiation.

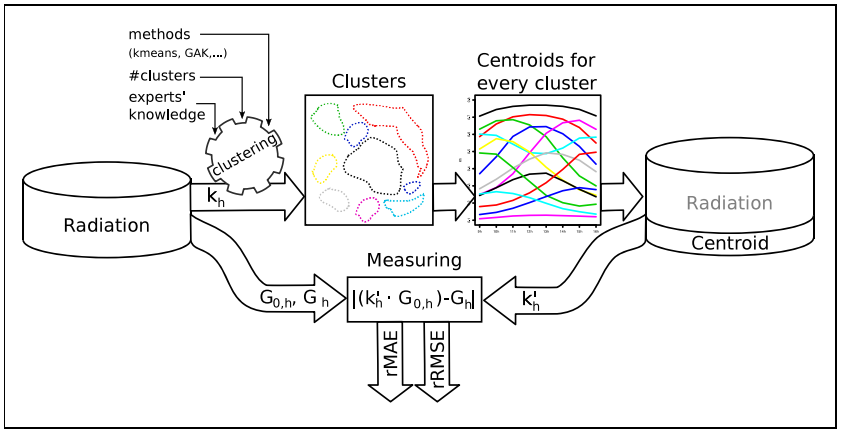

Fig. 2. First phase. Unsupervised learning with clustering methods. The first step determines the clusters (types of days), second step calculates the representative element for every cluster (centroids) and the third step evaluates the solution.

G_h) are used to test the validity of the proposed solution. Taking values of k_h into consideration, in the first step, clustering methods grouped examples in clusters showing similarities inside the same cluster and separation from others. Each cluster corresponds to a type of day. In the second step, the centroids for each cluster are calculated to be the representative pattern of that type of day (k'_h). The calculation is performed by averaging the values of all examples in each cluster.

Finally, as the last step in this phase, the solution needs to be evaluated and errors are therefore measured. As the values of $G_{0,h}$ and G_h are known, the deviations in estimated values (k'_h) will be detected by error metrics (rMAE and rRMSE).

Once the experts have validated the separation of days in different types, from a comprehensive point of view and supported by quality results in the measuring step, those centroids (types of days) are used to label the class attribute of the dataset. The augmented dataset then passes to the next phase.

3.2. Second phase (classification)

The second phase seeks to predict the hourly global solar radiation for next day with information available until the current day: meteorological (observed during present day and prediction for next day) and atmospherics (radiation G_d and k_d observed during present day). Furthermore, the next day prediction itself (*real centroid*) that has been calculated in the first phase is now available in the training set. Fig. 3 shows a schematic diagram of the process in this phase.

Experts highlight the relevance of including the daily clearness index for next day (k_{d+1}). Some experiments using real data reveal that performance improves. However, the value of that variable for next day is not available during the present day and should be estimated (k'_{d+1}). Therefore, regression model (based on random forest) is trained with data enumerated above to calculate that estimation. The importance of including it is also revealed in the models induced by machine learning algorithms (see Section 5.2).

Yet again, different algorithms generate several models and the final system will be configured using the best generalizing model. Accuracy

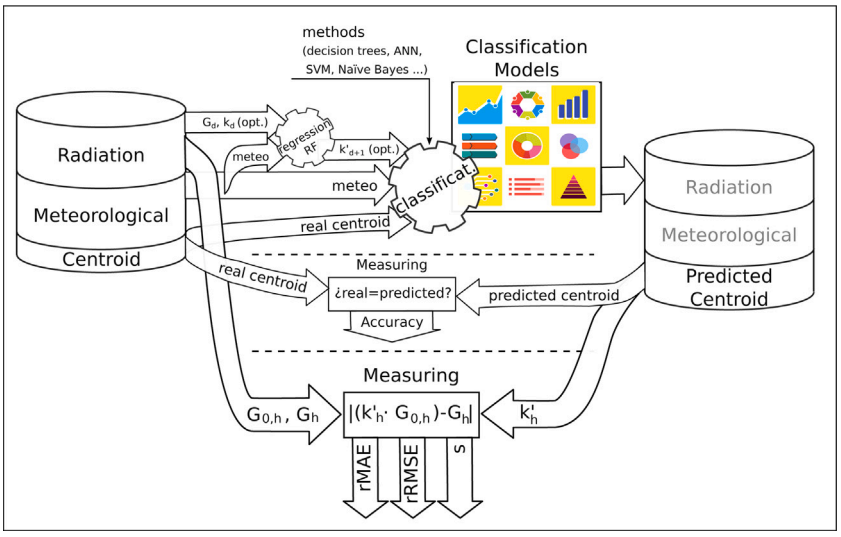

Fig. 3. Second phase. Supervised learning with classification methods. The first step induces the model with a previous optional step to estimate the daily clearness index for next day (k'_{d+1}). The second step calculates the internal accuracy and the third step evaluates the solution in terms of energy error (the most important criteria).

metrics can be measured to identify that model, but most important in this case is the final error observed (rMAE and rRMSE) in the prediction of hourly solar radiation. That error is calculated using metrics in the same way as in the first phase.

3.3. Predictive system

The final system, which is ready to predict, will store the set of centroids calculated in Phase 1 (one for each type of day) and will incorporate the model that better generalizes (and produces lower level of errors) data determined in Phase 2. When the meteorological and atmospherics data, available at the present day, are entered in the system, it will then respond with the type of day (centroid) estimated for next day. The hourly global solar radiation estimation (G'_h) can then be calculated using the centroid and the values of hourly extraterrestrial radiation. Fig. 4 represents this process graphically.

4. Experimental design

This section describes both the data and the algorithms and software used.

4.1. Dataset

A dataset with meteorological parameters and predictions of meteorological variables was used. Data were recorded from January 2005 to December 2015 at 10 Spanish locations under different weather conditions. According to Husein and Chung (2019) the evaluation of solar irradiance forecasting methods with several locations leads to greater confidence in the results.

Data were recorded by the Spanish Meteorological Agency (Agencia Estatal de Meteorología, AEMET). Forecasts were also obtained from the AEMET. Table 1 summarizes the description of these locations and the main meteorological parameters.

Data was prepared (pre-processed) in order to remove missing or incorrect values. According to Zhang et al. (2003), data preparation is important because real-world data may be incomplete, noisy and inconsistent. In order to solve this problem, we included in the preparation process:

- Elimination of observations with missing values in one or more attributes.

Table 1
Geographical coordinates and daily mean values of meteorological parameters for the locations used, 2005–2015.

Location	Lat/Long	G_d (kWh/m^2)	T_d (°)	H_d (%)
Albacete	38.95/−1.86	4.86	15.4	64
Alicante	38.37/−0.49	4.80	18.1	62
Barcelona	41.39/2.20	4.40	17.5	69
Bilbao	43.30/−2.04	3.45	14.8	73
Madrid	40.45/−3.72	4.75	15.0	57
Málaga	36.72/−4.48	5.11	19.2	67
Murcia	37.79/−0.80	4.97	18.8	62
San Sebastián	43.31/−2.91	3.48	13.5	80
Santander	43.49/−3.80	3.67	14.7	79
Toledo	39.88/−4.05	4.93	16.0	57

- Detection and elimination of noise and inconsistent values. We here followed the recommendations of the World Meteorological Organization (WMO) (Zahumensky, 2004). When an inconsistent value for an attribute was detected, the observation was removed from the data. That is, these data were not replaced or filled by any imputed value as the data set was large enough.

After cleaning missing, noise and inconsistent values, data preparation continued by selecting relevant data and by estimating new data from the original ones.

Instead of using hourly or daily values of global radiation (G_h, G_d), clearness index values were obtained (k_h, k_d). As has been explained in Section 2.1, the use of this index allows the daily and seasonal trend observed in solar global radiation series to be removed.

Two different data sets were then prepared as input for the analysed models. The first one only included the following exogenous measured variables: daily temperature and precipitation; in the case of humidity and temperature, the data set also included one value for periods from 9:00 to 12:00 and from 12:00 to 15:00; and the following predictions for the same 3-hours periods: temperature, humidity and cloudiness. The second dataset also included the daily clearness index estimated for next day (k'_{d+1}) as experts suggested.

For the first phase of the process, described in Section 3.1, a total of 33k data were used while in the second phase of the process, described in Section 3.2, a total of 26k were used; this is due to the availability of meteorological parameters predictions.

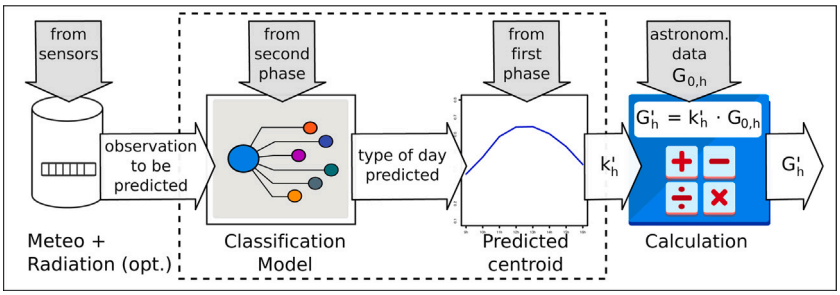

Fig. 4. Prediction phase. When meteorological and atmospherics data for present day are available, classifier can predict the type of next day and, therefore, the hourly clearness index for next day. The hourly global solar radiation (G'_h) is then estimated with that predicted hourly clearness index (k'_h) and the known extraterrestrial hourly solar radiation ($G_{0,h}$).

4.2. Algorithms and software

A wide variety of strategies to perform clustering and classification tasks were tested when selecting the algorithms that best perform in the proposed two-phases schema. Many options, presented in Section 2, were used. Different quality results were produced and they are discussed in Section 5.

The pre-processing steps, the induction of models (supervised or unsupervised), the calculation of metrics and other computational tasks were executed mainly by using R (R Core Team, 2019), some of its packages (Hornik et al., 2009; Sarda-Espinosa, 2019), and Weka (Witten et al., 2016).

In the first phase (clustering), we considered clustering approaches such as partitional (k-means, PAM, CLARA), hierarchical, fuzzy, density-based and kernel-based. The number of clusters considered was between 2 and 15. Some distance functions calculated were Euclidean, Pearson and DTW. We validate the quality of the generated clusters using internal measures such as silhouette width or the Dunn index. Taking into account all the possible combinations, we can argue that it is not affordable in a simple way and the knowledge of experts was decisive to guide the search. Their expertise was even more important because the type of days (centroids) needed to make sense and almost every non-guided proposal introduced some disturbing aspects.

In the second phase (classification), we used one well-known algorithm for each kind of methods described in Section 2.3. They are available in Weka and they were executed with the default configuration. Specifically, the algorithms used were: ZeroR (majority class as a baseline method), NaiveBayes, J48 (decision tree), LMT (logistic model tree), MLP (multilayer perceptron), SMO (support vector machine) and RandomForest (multiple classifier system). Furthermore, we included the persistence method, using the current class to predict the next one (without any learning process), as a baseline option.

5. Results

This section sets out the main results using the methodology explained in Section 3. One of the most important contributions is the characterization of types of days as this result can be useful in a broad variety of situations, beyond the prediction task considered in this paper. We also explain some patterns and relations discovered when exploring the models induced by different algorithms. This knowledge is useful for experts, it allows them to shed new light to current knowledge and to continue to advance in their research. The results are then validated with two objectives: to determine an appropriate configuration of the system, and to compare this new proposal with previous models. It will be shown that final configuration achieves highly accurate results. Finally, we enumerate some material made public to emphasize the reproducibility of these results.

5.1. Types of days

The most crucial problem in this research may be to determine the different types of days that should be considered in the process. That definition has implications on the understanding of the problem itself and on the successive decisions.

There is no consensus in the literature about the number of types of days and their election is usually determined by the specific location in the study. In our case, by using data from different climatic conditions obtained from 10 different locations, we aimed to increase the variety of data and to obtain more general results that can be applied in a wide assortment of cases.

Performing an intensive computation of every alternative is not deemed convenient, given the huge amount of data (examples and attributes) and possible methods to perform the clustering task (apart from the multiple parameters that can be tuned). That search must be guided by in-depth knowledge of the problem and the available data. We explored a small subset of combinations and analysed the results, but even with that reduced subset of alternatives, we can conclude there is no homogeneity regarding the number of clusters to be used. That parameter is decisive in the methodology, as it is possible to continue once the number of clusters have been determined.

In view of the difficulty, expert knowledge becomes fundamental in the search for the number of types of days: focusing on the hourly clearness index (k_h) data and dividing the problem into smaller ones, thus alleviating the problems detected when the whole dataset is considered. Therefore, a division into 3 categories of days is proposed: days with high, medium and low global radiation. The areas under the curves (AUC) defined by the 8 values of hourly clearness index (k_h) are estimated (using Section 2.1) to determine that global radiation. Once the AUC values have been calculated for every example, the minimum, average and maximum values are 0.42, 4.00 and 6.26, respectively. The theoretical minimum and maximum would be 0.0 and 7.0, so there always exists a level of global radiation and there is no example that starts and ends the day with maximum radiation, as expected. Fig. 5(a) shows how the most common days are those with highest AUC values.

The separation of days into three categories, taking in consideration that there is only one dimension (the AUC value), is performed with k-means and with Jenk's natural break optimization (Jenks, 1977). Both methods provide the same solution: using two breaks in 2.59 and 4.28. Thus, three separate intervals are created: [0.42, 2.59), [2.59, 4.28) and [4.28, 6.26]. Fig. 5(b) shows the frequency of these three categories of days. The subset with highest AUC values comprises half of the examples. Once three categories of days have been identified, it is easier to conduct clustering identification in each interval.

Some other decisions help to focus the efforts to find a solution. For example, considering the type of data to be clustered (eight consecutive values of hourly clearness index k_h), the distance measure that best fits in the clustering methods is those based on shape, such as DTW. Specifically, in this context, TGA kernels (Cuturi, 2011) achieve high

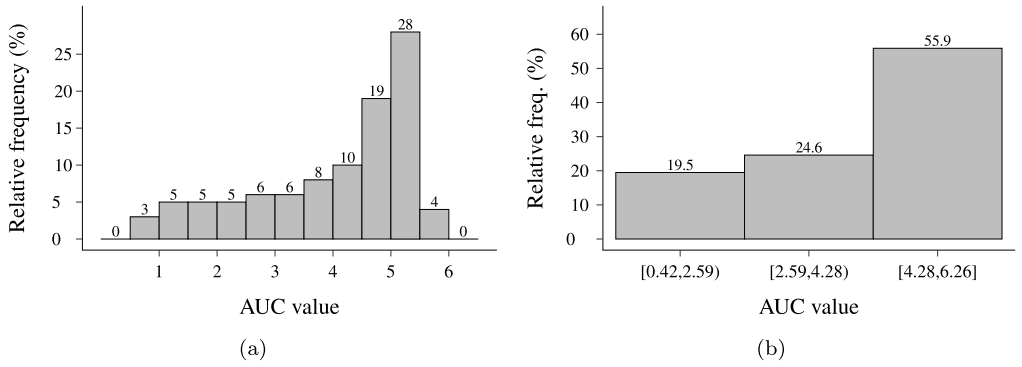

Fig. 5. (a) Histogram for AUC values and (b) histogram after dividing original data in three separate intervals.

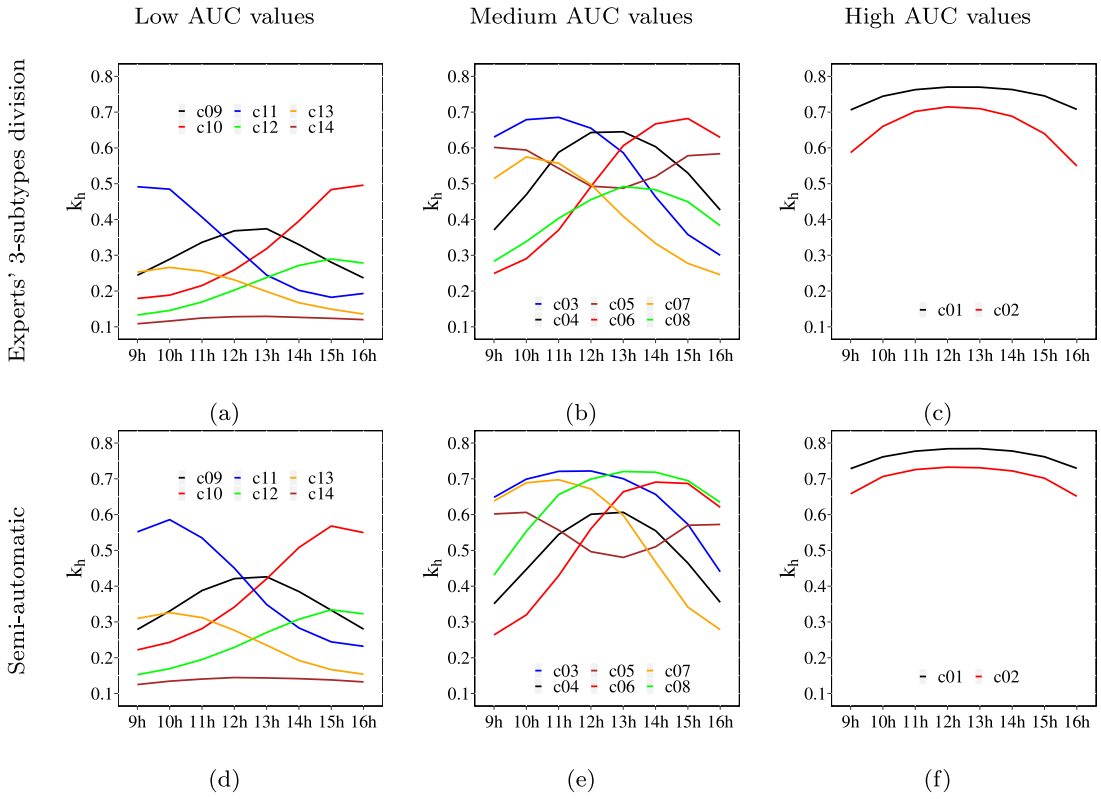

Fig. 6. Types of days. 14 types of days have been identified. The clustering in the upper row follows the experts' advice (using three categories for the day). In the lower row, the clustering uses a semi-automatic approach (there are no categories and graphical separation is made for better comparison).

quality results. Additionally, expert knowledge advises that a relatively small number of clusters for each category of day should be considered.

The final configuration that is proposed defines two clusters in the interval of highest AUC values and six clusters in the intervals of medium and lowest AUC values. Therefore, 14 types of days are being considered. In Fig. 6, the three upper plots represent those types of days (centroids) grouped by category of day.

Some important characteristics for the experts' decision to validate the result are: curves are symmetrical (in insolation or by pairs of two curves), they cover all the space, and they are sufficiently different from each other. For example, both curves with highest AUC values are symmetrical, they increase in the first half of the day and decrease in the second half of the day. In the case of the lowest AUC values, the centroids that obtain the highest values (lines green and red in upper chart on the left in Fig. 6) are symmetrical respectively: one (green line) starts at a high value and decrease while the other (red line) starts at a low value (similar to the end of the first one) and increase to a high value (similar to the starting of the first one). The error from summarizing all the examples by their corresponding centroids is low. Specifically, the rMAE is 10.7% and the rRMSE is 15.9%. This is another outstanding aspect that reveals that the selection seems to be appropriate.

Once the number of clusters were determined using expert advice (with 3 categories of days), we re-executed the clustering algorithm to search for the best 14 types of days (without any previous categorization). This experiment can be considered to be a semi-automatic search for the clusters as it only uses partial knowledge from the previous step (the number of clusters) but it does not need any other help from the experts. Errors calculated for this semi-automatic approach are similar to those from the previous result: the rMAE is 10.1% and the rRMSE is 15.2%. The three lower plots of Fig. 6 show the centroids for the clusters. The separation in categories (low, medium and high) is not real and is performed for an easy comparison with the option that actually uses those categories. Matching centroids determined by both criteria was visually performed by the experts in an easy process. It can be seen that the shapes are similar, although the centroids are

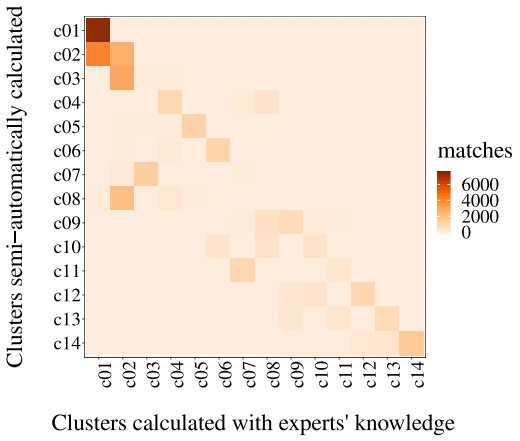

Fig. 7. Heatmap to represent the confusion matrix of clusters assigned using expert knowledge versus the clusters calculated semi-automatically.

slightly more overlapped in the semi-automatic version. That lower level of differentiation between centroids (clusters) can be responsible of a poorer performance in the second phase (classification) as we explain in Section 5.2.

Fig. 7 shows the confusion matrix of clusters determined by using the knowledge of experts (14 clusters from 3 categories of days) versus the clusters calculated semi-automatically (14 clusters directly). The heat map shows that most of the differences are between the assignment between clusters $c01$ and $c02$ (the two clusters with highest AUC values). However, those differences are not significant as both centroids are quite alike and have similar shapes.

The different types of days obtained with clustering, Fig. 6, correspond to different situations of the atmosphere according to the knowledge of experts; that is, the obtained clusters have a relationship with the observed real days. Thus, the shapes of days included in Fig. 6(a) correspond to days with low daily and, therefore hourly, clearness index; the ones included in Fig. 6(b) correspond to days with medium daily clearness index and Fig. 6(c) corresponds to days with high daily and hourly clearness indexes. The two first, low and medium daily clearness index, include six different shapes that correspond to the different hours at which hourly clearness indexes are low, that is, to the different moments when there can be clouds (as they are the most important radiation attenuation factor). There are different patterns that represent different types of day:

- Days with lower hourly clearness indexes in the morning, that correspond to situations in which there are clouds in the morning. These type of days are in red and green in Fig. 6(a) (c10 and c12) and in Fig. 6(b) (c06 and c08). The lower values of hourly clearness index (green line) can be explained by the presence of a type of clouds that absorb more radiation.
- Days with lower hourly clearness indexes in the evening. The explanation of experts is similar to the previous type. In this case, these type of days are in blue and orange in Fig. 6(a) (c11 and c13) and in Fig. 6(b) (c03 and c07).
- Days that always have very low values of hourly clearness index. This corresponds to completely overcast day, shown in brown in Fig. 6(a) (c14).
- Days with low values of clearness index specially in the morning and evening. This correspond to days in black in Fig. 6(a) (c09) and in Fig. 6(b) (c04).
- Days with low clearness index values at noon, such the day in brown in Fig. 6(b) (c05).
- Days with high clearness values for the whole day, as days in Fig. 6(c). Among these, there are two types, one that corresponds to days with a clean atmosphere and high incidence angle of solar radiation (black line – c01 –) and another that corresponds to days without clouds but with more attenuation of solar radiation due to the composition of the atmosphere (red line – c02 –).

The clusters obtained in a semi-automatic way (Figs. 6(d), 6(e) and 6(f)) are similar to the types of days described above.

5.2. Classification

The core of the second phase in the proposed methodology is the classification task. It takes the information of the type of days (centroids of the clusters) included in the original dataset and tries to learn which type of day will be next day (one-day ahead forecast) for different situations.

There are different approaches to carry out the classification task, as set out in Section 2.3. We selected very well-known algorithms for every approach in order to get accurate results, but, at the same time, some could provide understandable information. Models such as decision trees or logistic regression can be easily interpreted (Carvalho et al., 2019). They can be recognized as common alternatives in the context of eXplainable Artificial Intelligence (XAI).

The most accurate model in the classification phase and the one that minimizes errors (as presented in 5.2.2) is built by LMT (Logistic Model Trees) algorithm (Landwehr et al., 2005) when using the dataset that includes the daily clearness index estimated for the next day (k'_{d+1}). Their results are marked in italics in the tables with the results. However, the simplicity of the induced model is more interesting than the good quality achieved. The tree expands only once, the attribute selected at the root is the daily clearness index estimated for the next day creating two branches ($k'_{d+1} \leq 0.56$ and $k'_{d+1} > 0.56$). Other relevant attributes present in the logistic models at the leaf nodes are the prediction of humidity and cloudiness forecast. The importance of the daily clearness index estimated for next day (k'_{d+1}) is also detected in the tree induced by the J48 algorithm (attribute in root node and subsequent nodes). Therefore, the importance of including that estimation suggested by the experts is notable.

In addition to describing the most relevant patterns, the results regarding the accuracy of models with some level of confidence need to be given. Validation is needed in order to avoid optimistic results suffering from overfitting. The most common procedure in the literature related to analysing solar radiation is the division of the dataset into two subsets (Gutierrez-Corea et al., 2016; Zhang et al., 2017). Thus, the traditional validation, presented in 5.2.1, uses 70% of the dataset to train, and the remaining 30% for testing. On the other hand, from the field of machine learning, it is more usual to conduct a different procedure based on the repetition of 10-fold cross validations (James et al., 2013), which results are presented in 5.2.2.

5.2.1. Traditional validation

The first evaluation of the proposed model has been performed using the metrics described in Section 2.4 and dividing the dataset into two different subsets, one for training (70%) and one for testing (30%). The performance of the different analysed models is shown in Table 2.

In the results, only the second data set (the one that includes k'_{d+1}) has been showed because the results are better, as revealed by the comparison between the two data sets performed in Section 5.2.2. To forecast the next-day label , the classification accuracy of the models using expert knowledge clearly outperforms in terms of the correctly classified days that ranges from 9 to 10% compared to the models not using expert knowledge (semi-automatic).

The baselines models such as Persistence and ZeroR, represent the lowest accuracy while the LMT model achieves the maximum accuracy that is equal to 66.11% in the expert experiment compared to 57.22% for semi-automatic one, which is not so far from that of MLP, SMO and RF models.

The next step is based on the classification results, where the centroid matching with the predicted day is used to calculate a prediction

Table 2
Accuracy, relative Mean Absolute Error (rMAE) and relative Root Mean square error (rRMSE) achieved for different methods considering the dataset *with* global radiation information (so daily clearness index for the next day – k'_{d+1}– can be estimated). Best results are highlighted in bold. This evaluation considers a traditional validation, so models have been trained by using 70% of dataset and values given in this table are the result of testing with the remaining 30%.

	Accuracy (%)		rMAE (%)		rRMSE (%)	
	Autom.	Experts	Autom.	Experts	Autom.	Experts
ZeroR	23.54	36.45	30.07	28.69	46.45	44.87
Persist.	–	–	27.44	27.44	42.16	42.16
NB	39.65	48.51	15.45	15.65	23.89	24.17
J48	49.43	59.34	13.97	13.95	21.62	21.38
LMT	***57.22***	***66.11***	*12.71*	***12.86***	*19.81*	***19.82***
MLP	56.54	65.81	**12.67**	12.92	**19.74**	19.85
SMO	54.69	63.78	12.78	13.07	20.09	20.28
RF	56.15	65.44	12.94	13.09	20.24	20.20

Table 3
Accuracy achieved in classification with two different data sets using two different definitions of types of day. Datasets include (or not) global radiation information (to estimate daily clearness index for the next day – k'_{d+1}–). Definition of types of days consider two different criteria (using three categories of day – based on experts knowledge – or not using it – semi-automatic–). Mean and standard deviations values are given (calculated on a 10 × 10-fold cross validation). Best results are highlighted in bold. Reference model (LMT) is highlighted in italics.

Acc. (%)	Without k'_{d+1}		With k'_{d+1}	
	Semi-autom.	Experts	Semi-autom.	Experts
ZeroR	23.54 ± 0.02	36.46 ± 0.02	23.54 ± 0.02	36.46 ± 0.02
NB	29.66 ± 0.72	39.35 ± 0.69	39.52 ± 0.84	49.64 ± 0.69
J48	26.39 ± 0.81	36.11 ± 0.76	50.76 ± 0.82	60.88 ± 0.77
LMT	*32.56 ± 0.62*	*43.35 ± 0.71*	***58.26 ± 0.82***	***67.10 ± 0.74***
MLP	34.19 ± 0.81	43.85 ± 0.86	57.39 ± 0.98	66.70 ± 0.83
SMO	32.24 ± 0.65	42.87 ± 0.49	57.19 ± 0.77	66.10 ± 0.63
RF	**34.47 ± 0.74**	**44.04 ± 0.61**	58.00 ± 0.70	67.12 ± 0.64

of one day ahead hourly global radiation (G'_h). This prediction is compared to the measured data (G_h). The rMAE and rRMSE are also presented in Table 2.

The baseline models (Persistence and ZeroR) obtain the highest levels of errors as expected. However, it is noticeable that the algorithms with the greatest accuracy do not necessarily lead to the same algorithms in terms of the error metrics to estimate hourly global radiation. For example, NB and J48 versus MLP model. This is due to the difference in level of the height of the irradiance distribution between the predicted and real data in the misclassified part. For the rest of the models, we can also note that many results are similar, where the best rMAE and rMSE (shared between LMT, MLP, SMO and RF) are around 12% and 20% respectively, keeping in mind that the models that incorporate the expert knowledge about the type of days show a similar performance to those that selected semi-automatically those type of days.

However, it is not possible to confirm whether the similarities observed are casual or due to the specific division that has been made between the test and training set. A statistical validation is performed in the next subsection to confirm that the simulation is well established and it is independent of the specific results of the division of the data.

5.2.2. Statistical validation

We have two different datasets (with or without daily clearness index estimation for the next day –k'_{d+1}–) labelled with the type of day (learnt in the clustering phase) considering two different criteria (using three categories of day – based on experts knowledge – or not using it – semi-automatic–).

We repeated 10 times a 10-fold cross validation to validate the results and obtain an estimation of the generalization capacity of the induced models. That means that 100 experiments were executed for every combination. The Wilcoxon's statistical test was then used to find differences between every algorithm with respect to the selected model, LMT. That information is represented in the tables with the $\oplus$ and $\ominus$ symbols. This statistical validation is useful when values between alternatives are slightly different, because it determines with a confidence level (in this case $\delta = 0.05$) whether they can be considered as different.

Knowing how the cluster assignment in the first phase is learned is of interest before moving on to evaluate the second phase. This can be measured by using the accuracy. Table 3 presents the accuracy values (mean and standard deviation) for the two datasets considered using semi-automatic and experts-based approaches. It can be seen that including information about k'_{d+1} produces a preeminent improvement in the prediction of next day type of day. That information is very important, as revealed in the pattern used by the explainable models (LMT or J48). There is another improvement in the accuracy when using the types of days divided in 3 categories (experts' proposal) versus the semi-automatic approach (with none categorization). It can be motivated by a greater degree of overlapping between centroids.

Accuracy is not decisive to identify the best configuration for the second phase, because the most important measure is the error in the estimated radiation at the terrestrial surface ($G'_h - G_h$). That error is similar when using 3 categories for the type of day (experts-based approach) and when they are not used (semi-automatic approach), despite the better performance in the classification of the experts-based approach. In the following tables (see Tables 4 and 5) only the results for the experts-based approach are presented. Two datasets (with or without k'_{d+1}) are used, and once again, the best results are observed in the dataset with daily clearness index estimation for next day (k'_{d+1}). The results achieved when it is not used (and only meteorological data are available) are comparable to some previous works (see 5.3). However, a very interesting point is that the methodology proposed to generate the model is very simple and it only uses data that can be obtained from Meteorological Agencies. Additionally, if the user has access to the global radiation (radiation G_d and k_d observed during present day), the prediction will be much more precise, because the daily clearness index for next day is estimated (k'_{d+1}) and it is available for the most accurate system.

Majority class (ZeroR) and persistence models are selected as baseline classifiers. Every classifier outperforms those baseline models, so there is information in the dataset that is being used to improve the classification. In fact, the improvement, measured by s index is close to 50% when global radiation information is considered (and k'_{d+1} is estimated). We argue that, besides global radiation, there are some other attributes that are usually present in the patterns discovered, such as humidity or cloudiness predictions for next day.

Given the results in Table 5, we can see that there are significant differences from the statistical point of view. However, differences between those results are close: less than 3% in rMAE and less than 5% in rRMSE. It seems that the selection of the classification algorithm is not as important as the definition of the methodology. With this two-phases proposal, accurate results are reached in the second phase thanks to the simple methodology and the information generated in the first phase.

Although different classification algorithms get close results, we have selected LMT as the reference model, because of its better performance and its explainability. In addition, this system will be very fast during the prediction phase as it only needs to calculate 14 logistic regressions. Thus, its complexity for prediction is constant too, regardless of the observations.

5.3. Validation versus previous models in the literature

We have checked the results of the proposed methodology against several previously proposed models. One of the main problems when

Table 4
Relative Mean Absolute Error (rMAE), relative Root Mean square error (rRMSE) and forecast skill over 24 h persistence forecast (s) achieved for different methods considering the dataset *without* information about global radiation (so k'_{d+1} is *not* estimated). Mean and standard deviation values are given(calculated on a 10 × 10-fold cross validation). Best results are highlighted in bold. Reference model (LMT) is highlighted in italics. Symbol ⊕ (or ⊖) shows that such model is statistically better (or worse) than reference model (LMT) based on paired samples Wilcoxon test with $\delta = 0.05$.

	Without k'_{d+1}			
	rMAE	rRMSE	s	
ZeroR	28.66 ± 0.22	44.78 ± 0.4	−6.2 ± 1.37	⊖
Persist.	27.44 ± 0.41	42.15 ± 0.58		
NB	20.88 ± 0.28	32.66 ± 0.45	22.5 ± 1.56	⊕
J48	22.41 ± 0.34	34.25 ± 0.5	18.7 ± 1.72	⊖
LMT	*20.29 ± 0.29*	*32.83 ± 0.51*	*22.1 ± 1.59*	
MLP	20.02 ± 0.37	32.35 ± 0.67	23.3 ± 1.9	⊕
SMO	20.82 ± 0.26	33.88 ± 0.48	19.6 ± 1.37	⊖
RF	**19.77 ± 0.23**	**31.94 ± 0.4**	**24.2 ± 1.45**	⊕

Table 5
Relative Mean Absolute Error (rMAE), relative Root Mean square error (rRMSE) and forecast skill over 24 h persistence forecast (s) achieved for different methods considering the dataset *with* information about global radiation (so k'_{d+1} *is* estimated). Mean and standard deviation values are given (calculated on a 10 × 10-fold cross validation). Best results are highlighted in bold. Reference model (LMT) is highlighted in italics. Symbol ⊕ (or ⊖) shows that such model is statistically better (or worse) than reference model (LMT) based on paired samples Wilcoxon test with $\delta = 0.05$.

	With k'_{d+1}			
	rMAE	rRMSE	s	
ZeroR	28.66 ± 0.22	44.78 ± 0.40	−6.2 ± 1.37	⊖
Persist.	27.44 ± 0.41	42.15 ± 0.58		
NB	15.39 ± 0.16	23.67 ± 0.29	43.8 ± 0.91	⊖
J48	13.73 ± 0.15	21.04 ± 0.26	50.1 ± 0.93	⊖
LMT	***12.81* ± 0.15**	***19.64* ± 0.27**	***53.4* ± 0.98**	
MLP	12.89 ± 0.19	19.77 ± 0.37	53.1 ± 1.10	⊖
SMO	**12.81 ± 0.13**	19.71 ± 0.25	53.2 ± 0.92	⊖
RF	12.88 ± 0.13	19.77 ± 0.25	53.1 ± 0.78	⊖

comparing results is that they do not all use the same metrics. On the one hand, we compare our results with those that use relative RMSE and MAE and, on the other hand, we compare our results with those that use RMSE and MAE. The main problem of this last comparison is that it depends on the total radiation for an hour, but those data are not available in the results presented in previous papers.

Regarding methods using relative metrics, Ghimire et al. (2019) propose the use of an half-hourly time-step predictive model (CLSTM) which integrates deep learning Convolutional Neural Network (CNN) with Long Short-Term Memory Network (LSTM). The model has been checked for one location in Australia. The reported rRMSE is 18.01% for 1-day ahead.

Regarding methods that use RMSE and MAE, Qing and Niu (2018) propose the use of long short-term memory (LSTM) networks and weather forecasting data for hourly day-ahead prediction (11 h a day). The model was trained and checked with a dataset of island of Santiago, Cape Verde. The RMSE obtained is 122.7 W/m_2. Husein and Chung (2019) propose the use of a deep long short-term memory recurrent neural network (LSTM) and compare their results with those obtained when a feedforward neural network (FFNN) is used, as this last method has proven to be useful in solar irradiance forecasting. Both models are used for data from 6 different locations.

Table 6 shows a comparative of the results obtained from different methods including the explained in Section 1.2.

5.4. Reproducibility and released software

To contribute to the reuse of the results presented in this contribution, we provide some data and code that implements several aspects of the methodology. In https://github.com/ursusdm, in the repository called predictingHourlySolarRadiation, we distribute: (a) description (values) of the 14 centroids used to characterized 14 types of days, (b) demo training dataset and code to conduct the first and second phases of the methodology, and (c) code to perform the prediction phase, configured with the best models determined in Sections 5.1 and 5.2.2.

Table 6
Results obtained from different approaches. Values marked with (*) are estimated with the data of the paper. N: number of locations. Sources: LSTM(1) and FFNN(1) from (Husein & Chung, 2019), LSTM(2) and BPNN(2) from (Qing & Niu, 2018) (test sets), ANN(3) from (Marquez & Coimbra, 2011), and MLP(4) from (Voyant et al., 2013).

Model	N	$RMSE/rRMSE$ (Wh/m^2/%)	$MAE/rMAE$ (%)	s
LSTM(1)	6	60.31–108.52/–	36.90–64.36/–	44.24–68.89
FFNN(1)	6	84.54–108.08/–	49.79-70.54/–	24.67-49.54
LSTM(2)	1	122.7/–	–	30.67(*)
BPNN(2)	1	150.3/–	–	17.84(*)
ANN(3)	1	–/20.0–23.0 (*)	–/–	–
MLP(4)	5	–/27.3–27.8	–/–	–
RF (without k'_{d+1})	10	157.3/31.9	97.3/19.8	24.2
RF (with k'_{d+1})	10	97.4/19.8	63.4/12.9	53.1
LMT (without k'_{d+1})	10	161.7/32.8	99.9/29.3	22.1
LMT (with k'_{d+1})	10	96.7/19.6	63.1/12.8	53.4

6. Conclusions

A simple methodology to perform one-day-ahead predictions of hourly global solar radiation is proposed. It includes the knowledge of experts that has been fundamental to develop the methodology. The knowledge generated in the process, mainly in the explainable models induced in the classification phase, could be very useful for experts to gain additional knowledge in their domain of expertise. The obtained predictions are necessary, among other applications, for decision making in the energy market in order to ensure the correct integration of grid-connected photovoltaic solar systems in power grid.

A highly appreciated contribution in the context of short-term prediction for solar radiation domain is the establishment of a categorization of types of days. There is no consensus about the characteristics (attributes) that should be used, nor the number of alternatives, nor the criteria to be considered. This proposal is based on a first identification of the type of days, and the good performance achieved let us deduce that it is a promising possibility.

The proposed methodology is simpler than previous alternatives. It uses only two phases: a first one that is able to capture some features and create new information (using clustering) and a second one that uses that new information to induce models (classification) based on the relationship between meteorological and radiation data (input) and the 1-day ahead global solar prediction (output). The extensive variety of days used to set up the model (11 years from 10 locations with different types of weather) is another positive aspect that has led to an accurate prediction system. Information about global radiation in the present day is needed in order to obtain best results, but even without that information, the results are comparable to more sophisticated alternatives that effectively need such global radiation information. Moreover, the proposed models can be recognized as common alternatives in the context of eXplainable Artificial Intelligence.

Regarding the errors, the relative root mean square error of the prediction model is less than 20%, which means a significant reduction on previous proposed models.

New products and services can be developed using the advances presented in this paper. For example, it will be easy to assess about the characteristics of new facilities of photovoltaic systems depending on the location or climate.

From the perspective of the methodology itself and the models induced in the classification phase, we seek to study the effect that fuzzy approaches could have. Consecutive binary divisions made to numerical attributes to create intervals could be compressed, which would allow us increase the comprehensibility of models, while improving accuracy.

One of the limitations of the proposed models is that they have been estimated and checked for locations whose latitudes range from $36°N$ and $44°N$, and mean global daily radiation range from 3.5 and 5.1 kWh/m^2. The data used are for locations with continental and Mediterranean climate. As future research, in order to generalize the proposed models, it would be desirable to check if these models are also valid for greater or lower latitudes where the distribution of type of days is different. One possible extension of the work could be to use the proposed methodology but refitting the models for regions with very different meteorological conditions.

CRediT authorship contribution statement

José del Campo-Ávila: Conceptualization, Data curation, Methodology, Software, Formal analysis, Investigation, Visualization, Writing - original draft. **Abdelatif Takilalte:** Software, Writing - original draft. **Albert Bifet:** Conceptualization, Writing - review & editing. **Llanos Mora-López:** Conceptualization, Data curation, Funding acquisition, 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

This work has been supported by the project RTI2018-095097-B-I00 at the 2018 call for I+D+i Project of the Ministerio de Ciencia, Innovación y Universidades, Spain.

References

Aghabozorgi, S., Seyed Shirkhorshidi, A., & Ying Wah, T. (2015). Time-series clustering - a decade review. *Information Systems*, *53*, 16–38.

Ayala-Gilardón, A., de Cardona, M. S., & Mora-López, L. (2018). Influence of time resolution in the estimation of self-consumption and self-sufficiency of photovoltaic facilities. *Applied Energy*, *229*, 990–997.

Bektas, E. (2014). A least squares support vector machine model for prediction of the next day solar insolation for effective use of PV systems. *Measurement*, *50*, 255–262.

Blaga, R., Sabadus, A., Stefu, N., Dughir, C., Paulescu, M., & Badescu, V. (2019). A current perspective on the accuracy of incoming solar energy forecasting. *Progress in Energy and Combustion Science*, *70*, 119–144.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832.

Chen, C., Duan, S., Cai, T., & Liu, B. (2011). Online 24-h solar power forecasting based on weather type classification using artificial neural network. *Solar Energy*, *85*(11), 2856–2870.

Coimbra, C. F., Kleissl, J., & Marquez, R. (2013). Overview of solar-forecasting methods and a metric for accuracy evaluation. In *Solar energy forecasting and resource assessment* (pp. 171–194). Elsevier.

Cuturi, M. (2011). Fast global alignment kernels. In *Proceedings of the 28th international conference on machine learning* (pp. 929–936).

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*(Jan), 1–30.

Deo, R., Wen, X., & Qi, F. (2016). A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. *Applied Energy*, *168*, 568–593.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, *29*(2–3), 103–130.

Frank, E., Wang, Y., Inglis, S., Holmes, G., & Witten, I. H. (1998). Using model trees for classification. *Machine Learning*, *32*(1), 63–76.

Gairaa, K., Khellaf, A., Messlem, Y., & Chellali, F. (2016). Estimation of the daily global solar radiation based on Box-Jenkins and ANN models: A combined approach. *Renewable & Sustainable Energy Reviews*, *57*, 238–249.

Ghimire, S., Deo, R. C., Raj, N., & Mi, J. (2019). Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms. *Applied Energy*, *253*, Article 113541.

Gutierrez-Corea, F. V., Manso-Callejo, M. A., Moreno-Regidor, M. P., & Manrique-Sancho, M. T. (2016). Forecasting short-term solar irradiance based on artificial neural networks and data from neighboring meteorological stations. *Solar Energy*, *134*, 119–131.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction* (2nd ed.). Springer.

Hornik, K., Buchta, C., & Zeileis, A. (2009). Open-source machine learning: {R} Meets {Weka}. *Computational Statistics*, *24*(2), 225–232.

Husein, M., & Chung, I. (2019). Day-ahead solar irradiance forecasting for microgrids using a long short-term memory recurrent neural network: A deep learning approach. *Energies*, *12*(10), 1856.

Iqbal, M. (1983). *An introduction to solar radiation*. New York, London: Academic Press, Inc.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R*. Springer.

Jenks, G. F. (1977). *Occasional paper*, *Optimal data classification for choropleth maps*. University of Kansas.

Jiménez-Pérez, P. F., & Mora-López, L. (2016). Modeling and forecasting hourly global solar radiation using clustering and classification techniques. *Solar Energy*, *135*, 682–691.

Kisi, O. (2014). Modeling solar radiation of mediterranean region in Turkey by using fuzzy genetic approach. *Energy*, *64*, 429–436.

Krakovsky, Y., & Luzgin, A. (2018). Robust interval forecasting algorithm based on a probabilistic cluster model. *Journal of Statistical Computation and Simulation*, *88*, 2309–2324.

Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. *Machine Learning*, *59*(1–2), 161–205.

Liu, Q., & Wu, Y. (2012). Supervised learning. In *Encyclopedia of the sciences of learning* (pp. 3243–3245). Boston, MA: Springer US.

Marquez, R., & Coimbra, C. F. (2011). Forecasting of global and direct solar irradiance using stochastic learning methods, ground experiments and the NWS database. *Solar Energy*, *85*(5), 746–756.

Mellit, A., & Pavan, A. M. (2010). A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at Trieste, Italy. *Solar Energy*, *84*(5), 807–821.

Monjoly, S., André, M., Calif, R., & Soubdhan, T. (2019). Forecast horizon and solar variability influences on the performances of multiscale hybrid forecast model. *Energies*, *12*(2264).

Muselli, M., Poggi, P., Notton, G., & Louche, A. (2000). Classification of typical meteorological days from global irradiation records and comparison between two Mediterranean coastal sites in Corsica Island. *Energy Conversion and Management*, *41*(10), 1043–1063.

NCSS Data Analysis Software Manuals (2019). Chapter 390: Area under curve. In *NCSS data analysis software manuals* (pp. 1–6).

Ozgoren, M., Bilgili, M., & Sahin, B. (2012). Estimation of global solar radiation using ANN over Turkey. *Expert Systems with Applications*, *39*, 5043–5051.

Perez-Ortiz, M., Gutierrez, P. A., Tino, P., Casanova-Mateo, C., & Salcedo-Sanz, S. (2018). A mixture of experts model for predicting persistent weather patterns. In *2018 international joint conference on neural networks* (pp. 1–8). IEEE.

Pierro, M., Felice, M. D., Maggioni, E., Moser, D., Perotto, A., Spada, F., & Cornaro, C. (2017). Data-driven upscaling methods for regional photovoltaic power estimation and forecast using satellite and numerical wheater prediction data. *Solar Energy*, *158*, 1026–1038.

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, C. Burges, & A. Smola (Eds.), *Advances in Kernel methods - support vector learning*. MIT Press.

Ponsard, C., Touzani, M., & Majchrowski, A. (2017). Combining process guidance and industrial feedback for successfully deploying big data projects. *Open Journal of Big Data*, *3*(1), 26–41.

Popescu, M.-C., Balas, V. E., Perescu-Popescu, L., & Mastorakis, N. (2009). Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, *8*(7), 579–588.

Qing, X., & Niu, Y. (2018). Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. *Energy*, *148*, 461–468.

Quinlan, J. R. J. R., & Ross, J. (1993). *C4.5 : programs for machine learning* (p. 302). Morgan Kaufmann Publishers.

R Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rana, M., & Rahman, A. (2020). Multiple steps ahead solar photovoltaic power forecasting based on univariate machine learning models and data re-sampling. *Sustainable Energy, Grids and Networks*, *21*, Article 100286.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *26*(1), 43.

Sarda-Espinosa, A. (2019). dtwclust: Time series clustering along with optimizations for the dynamic time warping distance.

Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janovsky, T. A., & Kamaev, V. A. e. (2013). A survey of forecast error measures. *World Applied Sciences Journal*, *24*(24), 171–176.

Voyant, C., Paoli, C., Muselli, M., & Nivet, M.-L. (2013). Multi-horizon solar radiation forecasting for Mediterranean locations using time series models. *Renewable & Sustainable Energy Reviews*, *28*, 44–52.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining, fourth edition: Practical machine learning tools and techniques* (4th ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc..

Wyner, A. J., Olson, M., Bleich, J., & Mease, D. (2017). Explaining the success of adaboost and random forests as interpolating classifiers. *Journal of Machine Learning Research*, *18*, 1–33.

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*(2), 165–193.

Zahumensky, I. (2004). *Guidelines on quality control procedures for data from automatic weather stations: Technical report*, (p. 9). Commission for basic systms. Expert team on requirements for data fro automatic weather stations. World Meteorological Organization.

Zhang, C., Zhang, S., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, *17*(5–6), 375–381.

Zhang, J., Zhao, L., Deng, S., Xu, W., & Zhang, Y. (2017). A critical review of the models used to estimate solar radiation. *Renewable & Sustainable Energy Reviews*, *70*(November 2016), 314–329.