UNIVERSIDAD
DE MÁLAGA

Dpto. Lenguajes y Ciencias de la Computación

Tesis Doctoral

Data mining models for short-term solar radiation prediction and forecast-based assessment of photovoltaic facilities

Autor:
Pedro Francisco Jiménez Pérez

Directora:
Llanos Mora López

2016

AUTOR: Pedro Francisco Jiménez Pérez

http://orcid.org/0000-0001-7858-4278

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga

La Dra. Llanos Mora López, Titular de Universidad en el Departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga

Certifica

Que Pedro Francisco Jiménez Pérez, Ingeniero en Informática, ha realizado bajo su dirección la tesis doctoral titulada Data mining models for short-term solar radiation prediction and forecast-based assessment of photovoltaic facilities, que se recoge en la presente memoria, cumpliendo todos los requisitos legales para optar al grado de Doctor, por lo que autoriza su lectura y defensa pública.

Y para que así conste y tenga los efectos oportunos, firmo este certificado en

Málaga, a 9 de junio de 2016

Dra. Llanos Mora López

A mis padres y a Priscill

Este trabajo de investigación ha sido financiado por la Junta de Andalucía (proyecto No. P10-TIC-6441).

Agradecimientos

Obtener el título de doctor es bastante gratificante aunque también es un proceso largo y tedioso que no sería posible sobrellevar sin la ayuda y compresión de toda la gente que te rodea. Es por eso que las primeras palabras de esta tesis doctoral son de agradecimiento a todas esas personas que han tenido algo que ver en el desarrollo de esta larga aventura...

Mi primer agradecimiento es para Llanos, de la que destacan su gran conocimiento, su rigor científico, sus ganas de mejorar cada aspecto de la investigación, su gran disponibilidad y dedicación. Sin ella esta tesis no hubiera sido posible y estoy seguro de que esta tesis es la mejor posible gracias a ella.

También agradecer a Mariano su dedicación, su gran conocimiento de la materia y sus ideas que seguro hacen de esta tesis un orgullo para mi.

Un agradecimiento muy especial a Michel por sus compañerismo, sus ideas, sus aportaciones y su rigor científico. Hemos compartido reflexiones, conversaciones y discusiones acerca de ciencia, política, religión y también hemos compartido piso.

También quiero agradecer a mis compañeros de investigación por su ayuda y los buenos ratos que hemos pasado en el laboratorio: Pedro, Rafa y Cristina.

Un agradecimiento muy especial y de corazón para toda mi familia ya que gran parte de lo que soy y lo que hago es gracias a ellos: mi padre, mi madre y mis cuatro hermanos.

El agradecimiento más especial es para Priscill por ayudarme, soportame, comprenderme y por su apoyo en la traducción al inglés de muchas de las palabras y expresiones que componen esta tesis.

Contents

List of Figures

Table Index

Nomenclature

α	Solar elevation
δ	Declination
γ	parameter for the kernel function in SVM
$\hat{A}$	Estimated value of A
λ	Longitude
ω_0	Constant that determines the activation threshold
ω_h	Hour angle
ω_i	Weight of input i
ω_{sr}	Sunrise angle
ϕ	Latitude
d	day
E_0	Eccentricity factor
G_t	Average values of measured global horizontal irradiation for period t
$G_{0,d}$	Daily extraterrestrial global radiation on a horizontal surface
$G_{0,h}$	Hourly extraterrestrial global radiation on a horizontal surface
$G_{0,t}$	Extraterrestrial global horizontal radiation on period t
H	Humidity
h	Hour
I_{sc}	Solar constant
$k^*_{h,d}$	Daily-detrended hourly clearness index
k_d	Daily clearness index
k_h	Hourly clearness index
k_t	Clearness index for period t

MAE	Mean absolute error
MSE	Mean square error
P	Presure
$rMAE$	Relative mean absolute error
$RMSE$	Root mean square error
s	Forecast skill over 24 hour persistence forecast
T	temperature
x	Value of an input or of a neuron of previous layer in ANN
ANN	Artificial neural network
DT	Decision tree
SVM	Support vector machine
SVM-C	Support vector machine for classification
SVM-R	Support vector machine for regression

Chapter 1

Introduction

As society progresses, the energy needs grow, we have more electronic devices that consume electricity, we travel to more places and more distant, more sophisticated goods are manufactured and factories that produce them require more energy. General energy consumption increases, especially the consumption of electric energy, devices that have traditionally worked with other energy sources now are electrifying itself, like vehicles (that worked with oil), bicycles (which traditionally had pedals), kitchens (traditionally powered with gas), etc.

Energy is a strategic resource and countries seek to be energy independent since an energy dependent country has to pay a large bill for the energy consumed and therefore its development is slowed considerably.

Integrating energy generation in factories and urban environments is interesting because that is where the most energy is consumed and avoid wasting energy in transport infrastructure as well as saving. Renewable energies are often the easiest to integrate into these environments because they require less infrastructure and cause fewer problems with noise, dirt, pollution, etc.

Traditional energy sources have serious problems: they are limited, produce harmful pollution on health and the environment and someday (fossil fuels) they will run-out, so mankind is forced to find clean and renewable energy sources.

At the same time scientific and technical developments advance, society is moving towards a more acute awareness of his responsibility to the environment. This is one fundamental reason that today renewable energy is undergoing a huge expansion, which raises numerous challenges to the scientific and technical community.

Renewable energy is clean, inexhaustible, and its technology is developing quickly thanks to initiatives as the Kyoto protocol, global commitment against global climate change. Renewable energy sources include: wind power, hydropower, solar energy, biomass, biofuel, etc.

The motivation for searching and developing clean energy sources comes from several facts: the greenhouse gases effect over environment (temperature raising, rain and snow decrease), (Core Writing Team and , eds.), depletion of some energy sources like gas and oil, (Bentley, 2002), the oil price behaviour (Gori et al., 2007), the oil resources, (Kjärstad and Johnsson, 2009), (Owen et al., 2010) and (Hughes and Rudolph, 2011). In this scene, solar energy, and more specifically the photovoltaic solar energy, is beginning to be increasingly important in the energy mix.

1.1 Photovoltaic solar energy current situation

The number of solar energy plants has increased significantly in recent years, mainly due to the following factors: the need to use energy sources that contribute to reducing carbon emissions, establishing support policies to introduce this type of systems, improving the efficiency of these systems and the significant reduction in the price of all the components that make up those systems. As the number of these systems rises, there is an increasingly greater need to develop systems that enable these energy sources to be integrated with the traditional generation system.

Renewable energy continued growing in 2014 at the same time of global energy consumption and despite the dramatic decline in oil prices during the second half of the year (REN21, 2015). The biggest growth took place in the power sector and was dominated by three technologies: wind, solar photovoltaic and hydropower. Solar PV and wind power have suffered significant cost reductions that have played an important role in the increasing electrification of transportation and heating applications and highlighting the potential for these two technologies in the future. In many countries renewable energy is broadly competitive with conventional fuels, particularly in the power sector. In developing countries distributed renewable energy systems offer a great opportunity to speed up the transition to moderm energy services and to grant energy access, although prices and financing are still being one of the major barriers. Renewables have become a mainstream energy resource, in some cases with a rapid expansion, contributing to diversification of the energy mix, however, growth in renewables capacity and improvements in energy efficiency are below the necessary rates to achieve the Sustainable Energy for All (SE4ALL) goals of doubling the renewable energy levels and energy efficiency improvements in order to provide universal energy access by 2030.

It's not clear what a distributed energy generation system is. Different definitions for Distributed Generation (DG) can be found in literature (Rújula et al., 2005), and some of then are not consistent, probably because they derive from different industries, experiences and objectives, but it's a widely used term.

Despite rising energy use, global carbon emissions associated with energy consumption remainded stable in 2014 due to the increased penetration of renewable energy and to improvements in energy efficiency. There is a global growing awareness that deployment of renewable energy is critical for addressing climate change, creating new economic opportunities and providing energy access to billions of people that still living without modern energy services.

Renewable energy provided an estimated 19.1% of the total energy consumed in 2013 and continued growing in 2014. Renewable energy support policies and decreasing costs are the main factors that have encouraged growth of renewable energy sources. At the same time, subsidies to fossil fuels and nuclear power are an important obstacle in developing countries. China was the first country in new renewable power capacity installations in 2014 while Brazil, India and South Africa led for the capacity added in their respective regions. Renewables accounted for 58.5% of net additions to global power capacity in 2014 and comprised an estimated 27.7% of the world's power generating capacity, enough to produce about 22.8% of global electricity, see figure 1.1.

Falling costs have made unsubsidised solar PV-generated electricity cost-competitive with fossil fuels. In 2014, solar PV reached another important milestone with an estimated 40 GW installed for a total global capacity of about 177 GW. Figure 1.1 shows the evolution of the different renewable technologies over the last 10+ years, including solar PV. The major part of this new capacity was installed in China, Japan and United States. Latin America and African countries had a significant new capacity added while most EU markets continued to decline for the third consecutive year.

Also, 2014 saw significants improvements in energy storage systems accross all

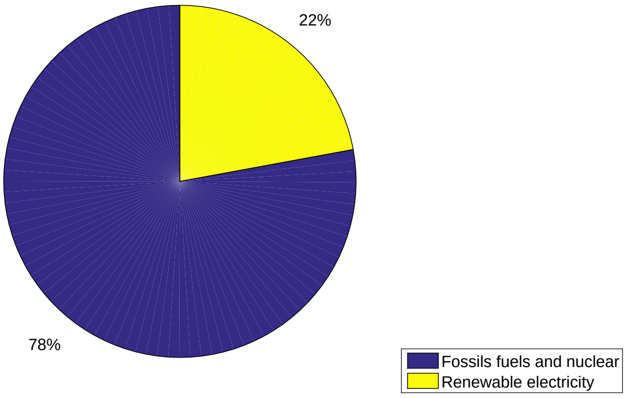


Figure 1.1: Global electricity supply in 2013, Source: GSR2015

sectors and several regions have seen significant growth in numbers of residential “prosumers” (electricity customers who produce their own power). In general, the solar PV industry recovery that began in 2013 continued in 2014, thanks to a strong global market.

According to data from (for Solar Energy Systems, 2015) photovoltaics is a fast growing market as the compound annual growth rate (CAGR) of PV installations was 44% from 2000 through 2014. China and Taiwan hold the lead of PV module production in 2014 with a share of 69% while Europe contributed with a 6% and USA and Canada each contributed with a 4%.

In 2014 Europe contributed with 48% of the total cumulative PV installations (in 2013 it was 58%) while China in Taiwan accounted for 17% (compared to 13% in 2013). 92% of the total production in 2014 was for Si-wafer PV technology. The share of multi-crystalline technology is now about 56% of total production and the share of all thin film technologies amounted to about 9%. The best cell efficiency is 25.6% for mono-crystalline and 20.8% for multi-crystalline silicon wafer-based technology while the best efficiency in thin film technology is 21.0% for CdTe and 20.5% for CIGS solar cells. In the last 10 years, the efficiency of average commercial wafer-based silicon modules increased from about 12% to 16% and CdTe module efficiency increased from 9% to 13%.

The Energy Payback Time of PV systems is dependent on the geographical location: while PV systems in northern Europe need around 2.5 years to balance

		START 2004	2013	2014
INVESTMET				
New investment in renewable power and fuels	Billion USD	45	232	270
POWER				
Renewable power capacity (total, not including hydro)	GW	85	560	657
Renewable power capacity (total, including hydro)	GW	800	1578	1712
Hydropower capacity (total)	GW	715	1018	1055
Bio-power capacity	GW	¡36	88	93
Bio-power generation	TWh	227	396	433
Geothermal power capacity	GW	8.9	12.1	12.8
Solar PV capacity (total)	GW	2.6	138	177
Concentrating solar thermal power (total)	GW	0.4	3.4	4.4
Wind power capacity (total)	GW	48	319	370
HEAT				
Solar hot water capacity (total)	GW_{th}	86	373	406
TRANSPORT				
Ethanol production (annual)	billion litres	28.5	87.8	94
Biodiesel production (annual)	billion litres	2.4	26.3	29.7

Table 1.1: Renewable power capacity until 2014

the input energy, PV systems in the south equal their energy input after 1.5 years and less, depending on the technology installed. For example, a PV system located in Sicily made with multi-Si modules has an Energy Payback Time of around one year, so, assuming 20 years of useful life, this system can produce twenty times the energy needed to produce it.

Inverter efficiency actually stands at 98% and above. The market share of inverters is estimated to be 50% for string inverters (residential, small and medium commercial applications), 48% for central inverters (with applications mostly in large commercial and utility-scale systems) and about 1.5% belongs to micro-inverters (used on the module level).

PV systems prices have reduced dramatically in last 20-25 years, for example, in Germany prices for a typical 10 to 100 kWp PV rooftop-system were around 14,000 €/kWp in 1990 while at the end of 2014, such systems cost about 1,300 €/kWp, showing a net-price regression of about 90% in that period of time (equivalent to an annual compound average price reduction rate of 9%). Nowadays China is the most important PV cell and module manufacturer, with about 60% of the total annual production in 2014. Europe and USA were big manufacturers 20 years ago but now they have a little share due to the extraordinary growth of China production.

Europe and Asia are the regions with most cumulative installed PV systems followed by America in third position. Germany accounts for the 20% of the world total, Italy accounts for the 10% and the rest of Europe for the 18%. Multi-Si is

Year	**GWh**	Annual change(%)
2011	255597	-1.9
2012	252014	-1.4
2013	246368	-2.2
2014	243544	-1.1
2015	**248181**	**1.9**

Table 1.2: Energy demand evolution 2011-2015

	MW	**MW %15/14**	**GWh**	**GWh %15/14**
Hydropower	18668	4.9	25733	-28.2
Nuclear	7866	0	56796	-1
Coal	10972	0	54553	23.8
Combined cycle	25348	0	26086	18.2
Rest Hydropower	2109	0	5659	-19.9
Wind	22845	0	47948	-5.3
Solar PV	4423	0.5	7861	0.8
Solar thermal	2300	0	5158	4

Table 1.3: Renewable installed capacity and power production in 2015

the most important technology produced nowadays, while Mono-Si stands for the second place. Thin-film is the less important. In 2014 Multi-Si accounted for 26.2 GWp of the total annual production, Mono-Si accounted for 16.9 GWp and Thin film 4.4 GWp.

1.1.1 The situation in Spain

In Spain, the most significant fact during the year 2015 was the rising demand for electricity after four consecutive years of decline (due to the crisis scenario), see table 1.2, although, renewable energy production suffered a descent due to decline in hydropower production. The peninsular installed power production capacity grew about 5% for hydropower and 0.5% for solar PV compared to year 2014, meeawhile others technologies remainded with little or no change. Solar PV accounted for 7861 GWh of electricity, a 0.8% increase over the year 2014, as seen in table 1.3.

In table 1.4 data about each technology production capacity share and the annual energy production share as percentages are shown. Note that some technologies capacity share can be about double of the final energy share it produced during the year and vice versa. Solar PV accounts for the 4.3% of the capacity share while in 2015 it produced 3.1% of the total produced energy. Combined cycle was the technology with the major installed capacity in 2015, with a share of 24.7% while nuclear had the highest produced energy share with a 21.7%. The

	capacity %	**produced %**	**MAX Hour**
Combined cycle	24.7	10.0	10.1
Coal	10.7	20.3	13.8
Nuclear	7.7	21.7	15.3
Cogeneration	6.9	10.6	8.5
Hydropower	20.2	11.1	14.5
Wind	22.3	19.1	35.3
Solar PV	4.3	3.1	0.1
Solar thermal	2.2	2.1	0.1
Geothermal	1.0	2.0	1.4

Table 1.4: Renewable installed capacity and power production in 2015

last column shows the share of produced energy for the maximum demand of the year 2015 in a one hour period, which was on february 4th between 20 a 21 hs. Since that hour of that month is on night, solar renewables sources had a poor share. The highest share was for wind with 35.3%.

Since 1998, the Spanish electricity market has moved from a centralized operational approach to a competitive one encouraging the deployment of solar plants with a financial penalty for incorrect prediction of solar yields for the next day on an hourly basis. In Spain, the called “Régimen Especial” (special regime) is suitable for power plants under 50 MW and which energy source is renewable energy. In this regime, power plants administrators can notify to the energy distributor firm with production forecast for each one of the 24 day periods 30 hours before the start of the indicated day and this forecasting is used to establish prices in the energy market. They can also make corrections over the previously notified forecasts until one hour before the day starts. If finally produced energy is different from forecasted energy in more than 5% in a day producton period (1 hour), the administrators will suffer an economic penalty for the deviation in energy production.

To stabilize demand and production, power sources must implement auxiliary services like balance between generation and consumption, voltage regulation and reactive power injection. Since 2010, photovoltaic systems with power over 2 MWp must contribute to the stability of the electric system against voltage gaps ((BOE, 2010)) applying some operational procedures that were previously applied to wind power sources (BOE, 2007).

1.2 Forecasting solar radiation and photovoltaic solar energy challenges

At the present time the photovoltaic solar energy faces some interesting challenges such as:

- Forecasting produced energy
- Energy demand management
- Use of battery systems

The first of these challenges will be addressed in this thesis.

It's easy to produce electricity from solar energy because is renewable, free and is available in almost every place on Earth and every day photovoltaic systems are cheaper and more efficient. Grid-connected photovoltaic power systems have to work collaboratively with other systems such as nuclear, gas, coal, hydroelectric or wind power plants, therefore, predicting energy production by these plants has thus become a requirement in competitive electricity markets. It is important to note that produced energy and consumed energy must be as close as posible everytime, so it's important to know how much energy power systems can produce in order to match power consumption. This is easier to achieve with traditional systems like coal or gas power plants, they can produce the desired energy everytime simply using more or less resources (fuel or so), but in other systems like hydropower or photovoltaic ones, they aren't able to produce the desired energy at each moment, they depent on the climatic conditions (water, solar irradiance) to produce energy. That's why it's necessary to develope energy production forecasting tools that let operate these kind of power plants in a safe and efficient way.

The energy distribution system stability must be guaranteed, that means, produced energy must be as closer as possible to consumed energy and forecasting techniques are necessary in the case of renewable energy source plants in order to know beforehand how much energy will they produce. In case of residential systems, knowing the amount of energy they will produce can be useful for the self-consumption improvement and get better profit.

The production of photovoltaic solar energy systems depends mainly on the availability of solar radiation that is variable and intermittent. Solar radiation varies due to weather conditions. The presence of clouds is one of the most important factors in terms of attenuation of this parameter. The problem is that the process of attenuation is highly stochastic making it difficult to predict the solar radiation.

The prediction of solar radiation can be considered as a problem of modeling and simulation and be addressed using data mining models. These models allow the analysis of large data sets and can be used to infer future behavior of them. One of the advantages of these models is their great generalization ability and the ability to work with different data types.

Related to solar energy forecasting, there are some issues that are currently interesting to research and improve, such as:

- Shor-term forecasting of solar radiation to forecast produced energy.
- Analyzing the possibility of combining several forecasting techniques into a unique model in order to improve results obtained with individual methods.

1.3 Objectives

The overall objective of this thesis is to develop data mining models to forecast solar global radiation 24 hours ahead and to use these predictions to evaluate the performance of photovoltaic systems.

This overall objective is specified in the following specific objectives:

1. Propose an index that allows us to remove the annual and daily trends observed in global hourly radiation data.
2. Analyze the probability distribution functions of detrened series.
3. Determine the minimum number of functions that allows us to capture all the information contained in the various distribution functions. For this specific goal, the proposal is to use a clustering technique and a statistical test to

check whether the selected probability distribution functions represent all observed probability distribution functions.

4. Analyze the different sources of meteorological variables that can be used to predict solar radiation.

5. Develop data mining models that allow including the different relationships observed between the radiation values of the next day depending on the values of the current day radiation and other meteorological parameters.

6. Develop the API to access external sources of meteorological data both measured and forecasted.

7. Development of a web system that include the proposed models to short-term forescast radiation.

8. Integrate the developed models in the evaluation models of photovoltaic solar systems.

1.4 Structure of this thesis

The thesis has been divided in eight chapters, including this Introduction, and four appendices.

Chapter 2 reviews other works about solar radiation forecasting, including some located in the mediterranean zone. Reviewed works include models based on statistical models, data mining models, satellite image based models, ground-based image models, numerical weather prediction models and mesoscale models.

Chapter 3 introduces the methods and models used in this work, like Cumulative Probability Distribution Function, used to characterize daily solar radiation profiles, Artificial Neural Networks and Support Vector Machines, as well as Linear Regression to predict contiuos values from past data. Also classification methods are presented like Decision Trees and Support Vector Machines for Classification. Finally performance metrics are presented to measure the accuracy of the proposed models. The data sets and data sources used in this work to test the proposed models are presented, including data from the meteorological station installed at University of Malaga, data from OpenWeatherMap website and data from AEMET (Agencia Estatal de Meteorología), the Spanish weather service.

Chapter 4 is dedicated to the solar radiation fundamentals, including astronomical concepts related to Earth-Sun position, characterization of solar radiation hourly series, clearnes index, used to remove seasonal trends, persistence model, used to compare with proposed models and the forecast skill, based on persistence model and used as reference model as well.

Chapter 5 introduces a model to forecast hourly solar global radiation using statistical methods like CPDF, K-means, and also using the clearness index presented in chapter 4. This models aims to predict the hourly solar radiation using the daily clearness index as input.

Chapter 6 details the proposed model to forecast the hourly global solar radiation using data mining methods and daily profiles of clearness index. K-means is again used to cluster daily solar radiation profiles, then a new variable is defined from the clearness index daily profiles. Support Vector Machines, Decision Trees and Artificial Neural Networks are used to predict the desired hourly solar radiation values.

Chapter 7 presents a methodology to assess solar power plants performance based on forecasted solar radiation. A OPC-based system is presented, which is able to obtain data from a large variety of equipment, then an algorithm to assess the performance of the plants using measured data and forecasted solar radiation data is developed.

Finally, chapter 8 is devoted to the conclusions of this thesis. It presents a summary of the topics discussed in this dissertation, taking into account the most relevant aspects of each chapter. Furthermore, prospective ideas for further research are described so as to ensure more studies in the solar radiation forecasting area.

Bibliography

Bentley, R., 2002. Global oil and gas depletion: an overview. Energy Policy 30 (3), 189 – 205.

BOE, 2007. Boletín oficial de estado. real decreto 661/2007, de 25 de mayo, por el que se regula la producción de energía eléctrica en régimen especial.

BOE, 2010. Bolet'ín oficial del estado. real decreto 1565/2010, de 19 de noviembre, por el que se regulan y modifican determinados aspectos relativos a la actividad de producción de energía eléctrica.

Core Writing Team, R. P., (eds.), L. M., 2014. Ipcc, 2014: Climate change 2014: Synthesis report. contribution of working groups i, ii and iii to the fifth assessment report of the intergovernmental panel on climate change. Tech. rep., IPCC, Geneva, Switzerland.

for Solar Energy Systems, F. I., 2015. Photovoltaics report. Tech. rep., Fraunhofer Institute for Solar Energy Systems, ISE.

Gori, F., Ludovisi, D., Cerritelli, P., 2007. Forecast of oil price and consumption in the short term under three scenarios: Parabolic, linear and chaotic behaviour. Energy 32 (7), 1291 – 1296.
URL http://www.sciencedirect.com/science/article/pii/S0360544206001873

Hughes, L., Rudolph, J., 2011. Future world oil production: growth, plateau, or peak? Current Opinion in Environmental Sustainability 3 (4), 225 – 234, energy Systems.
URL http://www.sciencedirect.com/science/article/pii/S1877343511000509

Kjärstad, J., Johnsson, F., 2009. Resources and future supply of oil. Energy Policy 37 (2), 441 – 464.

URL http://www.sciencedirect.com/science/article/pii/S0301421508005259

Owen, N. A., Inderwildi, O. R., King, D. A., 2010. The status of conventional world oil reserves—hype or cause for concern? Energy Policy 38 (8), 4743 – 4749.
URL http://www.sciencedirect.com/science/article/pii/S0301421510001072

REN21, 2015. Renewables 2015 global status report key findings. Tech. rep.

Rújula, A. B., Amada, J. M., Bernal-Agustín, J., Loyo, J. Y., Navarro, J. D., 2005. Definitions for distributed generation: a revision. Proceedings of the International Conference on Renewable Energy and Power Quality 05.

Chapter 2

State of the Art

2.1 Introduction

This chapter reviews other works about solar radiation forecasting, including some located in the mediterranean zone. Performance forecasting of systems that use solar radiation as energy resource requires, on the one hand, using actual forecasted running conditions (meteorological parameters) and, on the other hand, using models that permit estimating the operation performance of these systems taking into account these conditions. Estimating the energy produced by solar plants is difficult mainly due to its dependence on meteorological variables, such as solar radiation and temperature, (Luque and Hegedus, 2002), (Chang, 2009). In fact, photovoltaic production prediction is mainly based on global solar irradiation forecasts.

The main problem of having predictions of solar radiation that these systems will receive is in the nature of the solar resource, since it is an intermittent resource (due to day-night succession), it has a seasonal component, due to changes of relative position between Sun and Earth and has a certain stochastic behavior. The behavior of this variable can change dramatically on different days, even on the same day, due to the stochastic nature.

Prediction models of solar radiation must be able to collect these trends and to reproduce the non-deterministic component of it. In the following sections some of the main approaches used to predict solar radiation are analyzed and described.

It also includes a specific section in which the various errors of the main models are shown to have a reference to compare the proposals made in this work with.

2.2 Solar Radiation Forecasting Models

Multiple methods like statistical and data mining techniques can be used to address the need to forecast solar radiation. First attempts in irradiance forecasting were made more than twenty years ago (Jensenius and Cotton, 1981), when daily solar radiation forecasts for one to two days in advance have been produced with the Model Output Statistics (MOS) technique (Glahn and Lowry, 1972). Subsequent years showed only minor attemps or progress with respect to the development of solar irradiance forecasting methods. (Heck and Takle, 1987) and (Jensenius, 1989) both presented variations of the MOS approach without introducing new elements.

After these first approaches, numerous works and approaches have been proposed to characterize and predict solar radiation. Some of these approaches are based on using physical models, others, conversely, are based on the assumption that the prediction of solar radiation can be tackled as others forecasting processes. Here is a review of studies about solar irradiance forecasting using different methods.

2.3 Statistical and data mining models

Statistical methods based on historical data can be divided in two categories: statistical and data mining models. Examples of statistical methods are: seasonality analysis, Auto Regressive Moving Average (ARMA). Examples of data mining models are fuzzy inference, genetic algorithm and neural networks.

2.3.1 Statistical models

Statistical time series models are based on the assumption that the data have an internal structure and it can be identified by using simple and partial autocorrelation, (Box and Jenkins, 1976), (Gooijer and Hyndman, 2005), (Brockwell and Davis, 2002). Time series forecasting methods detect and explore such a structure. In particular, ARMA (autoregressive moving average), ARIMA (autoregressive integrated moving average) models have been widely used; for instance, (Brinkworth, 1997), (Bartoli et al., 1983), (Aguiar et al., 1988), (Graham et al., 1988), (Aguiar and Collares-Pereira, 1992) and (Mora-López and de Cardona, 1998) propose different methods for modeling hourly and daily series of clearness index (parameter related to solar global radiation). These models are particularly useful for long-term characterization and prediction of the clearness index as they pick up the statistical and sequential properties of series. However these methods have not been used for short-term prediction of clearness index as the error in the prediction of isolated values (next value in a series) is too large.

(Mora-Lopez and Sidrach-de Cardona, 1998) proposes a methodology to generate hourly series of global irradiation. The only input parameter required is the monthly mean value of daily global irradiation, which is not difficult to be available for many locations. The procedure is based on multiplicative autoregressive moving-average (ARMA) statistical model for time series with regular and seasonal components and is able to capture the two relationships observed in recorded hourly series of global irradiation: the relationship between one value and the value of the previus hour and the relationship of the value and the value of the previous day at the same hour. Data from several Spanish cities and ranging from year 1976 to 1986 are used. Daily and seasonal trends are removed using the maximum hourly global irradiation value and a difference operator with hourly values from a given day and the prevoius day. In order to check the validity of the proposed methods to estimate the unknown parameters of the multiplicative ARMA models, synthetic hourly series of clearness index and global irradiation were obtained and it was verified that the generated series (for both parameters) have the same statistical characteristic as the real series, that is, same mean, variance and cumulative probability distribution function as the real series (using 0.05 as significance level in Kolmogorov-Smirnov test)

In (Reikard, 2009) six data sets (three from Kansas City (Missouri), Denver (Colorado), and Phoenix (Arizona), and the others from the Measurements and Instrumentation Data Center baseline measurement system database: Clark power station in Nevada, Solar Radiation Research Laboratory (SRRL) and The National Wind Technology Laboratory) are used to run forecasting experiments at resolutions of 5, 15, 30, and 60 min, using the global horizontal component. Forecasting tests are run using regressions in logs, Autoregressive Integrated Moving Average (ARIMA), and Unobserved Components models but also transfer functions, neural networks, and hybrid models are evaluated. The models are estimated over history prior to the start of the forecast horizon, the data is forecasted and then the predicted values are compared with the actuals. The best results are obtained using the ARIMA in logs in nearly all the tests, with time-varying coefficients, but with some exceptions: at high resolutions, a transfer function using cloud cover is found to improve over the ARIMA and in a few cases, the neural net or hybrid models can improve at very high resolutions (5 min). The ability to capture the diurnal cycle more effectively than other methods explains the success of the ARIMA method.

A new approach divided into two phases is used to predict the hourly solar radiation series in (Ji and Chee, 2011). The Autoregressive and Moving Average (ARMA) model is used to predict the stationary residual series, previously detrended selecting the best model based on the Augmented Dickey–Fuller method to test the stationarity of the residual. Furthermore, a hybrid model that combines both the ARMA and Time Delay Neural Network (TDNN), is applied to produce better prediction, where ARMA model is used to predict the linear component of the series and the TDNN model is used to predict the nonlinear component. The simulation shows that this hybrid model can take the advantages of both ARMA and TDNN and give better results than applying only the ARMA model.

As explained in (Diagne et al., 2013), statistical models based on online irradiance measurements can be used to forecast at very short term, from 5 minutes up to 6 hours. Some examples are Auto Regressive (AR) and Auto Regressive Moving Average (ARMA) models.

2.3.2 Data mining models

Data mining techniques have been also applied for process forecasting. These approaches have been proposed to overcome the limitations of statistical methods because they do not require any assumptions to be made, particularly with respect to the linearity of the series.

In (Sfetsos and Coonick, 2000) a new approach for the forecasting of mean hourly global solar radiation on a horizontal surface is introduced. Arfificial Neural Networks (ANN) and the Adaptative Neuro-fuzzy Inference Scheme (ANFIS) are used. Initially one variable is used and then experiment is extended to include aditional meteorological parameters. Results indicate better performance for this artifical intelligence model over conventional procedures based on clearness index and are able to capture de periodic nature of this series. Levenberg-

Marquardt (LM) network was found to be the optimal prediction model (against back-propagation). Use of aditional meteorological variables such as temperature, pressure, wind direction and speed as potential input for the forecasting process is researched; some models can be further enhanced using aditional meteorological parameters but not all. LM and ANFIS performance are enhanced when using wind direction parameter as input. The best prediction is found to be that from the multivariate LM case, with an RMS error improvement about 74% compared with that of the bench-mark persistent aproach. In this approach no transformations of the solar radiation values are required.

In (Perez et al., 2007) a simple solar forecast model using sky cover predictions is developed and tested against both ground-measured and satellite-derived irradiances data. Data is taken from the National Digital Forecast Database [NDFD] of United States National Weather Service, providing gridded forecasted parameters for the entire country. Three methods are tested for different forecasting periods, resulting that 'Best fit formula' is the best method with a Relative Mean Bias Error of -2% and a Relative Root Mean Square Error of 35% for less than 4 hours ahead forecasting.

Artificial neural network and ARIMA models are proposed in (Reikard, 2009); the errors range from 30 to 40% in energy terms. Similar models are used in (Voyant et al., 2013); the obtained errors for predicting hourly values for a day range from 23% to 28%.

(Mellit and Pavan, 2010) proposes a practical method for solar irradiance forecast using artificial neural network (ANN), a Multilayer Perceptron MLP-model that makes it possible to forecast the solar irradiance on a base of 24 hours using the present values of the mean daily solar irradiance and air temperature. A database of solar irradiance and air temperature data (from July 1st 2008 to May 23rd 2009 and from November 23rd 2009 to January 24th 2010) collected in Trieste, Italy, is used. A K-fold cross-validation was carried out in order to check the generalization capability of the MLP-forecaster. The results indicate a good performance, as the correlation coefficient is in the range 98–99% for sunny days and 94–96% for cloudy days. A comparison between the forecasted energy and the one produced by the GCPV plant installed on the rooftop of the municipality of Trieste shows the goodness of the proposed model. The MLP input layer accepts the mean daily solar irradiance, the mean daily air temperature and the day of the month (at the time t) as parameters, while the output layer gives the 24 h of solar irradiance at the next day (time t + 1). The results show that the developed MLP-forcaster is suitable for the prediction of sunny days with r values between 98% and 99%), and it also provides acceptable results for cloudy days where r is between 92% and 95%. For the whole dataset the MBE and the $RMSE$ are 32% and 67% respectively.

In (Mora-López et al., 2011) a model for short-term forecasting of continuous time series of solar radiation that binds the use of both statistical (regression) and machine learning methods has been performed. The mean square errors of the proposed models range from 0.04 to 0.4 depending on the value of clearness index.

An artificial neural network (ANN) model was used to estimate the solar radiation parameters for seven cities from the mediterranean region of Anatolia in Turkey in (Koca et al., 2011). The maximum $RMSE$ was found to be 6.9% for Mersin citation and the best value was obtained to be 3.6% for Isparta.

A medium-term solar irradiance forecasting model was developed in (Marquez and Coimbra, 2011) adopting predicted meteorological variables from the *US National Weather Service's (NWS)* forecasting database as inputs to an Artificial Neural Network (ANN) model. The inputs involved are the same from a validated forecasting model so mean bias error (MBE), root mean square error ($RMSE$) and correlation coefficient ($R2$) comparisons between the more established forecasting model and the proposed one are included. A set of criteria for selecting relevant inputs was developed, input variables were selected using a version of the Gamma test combined with a genetic algorithm. The solar geotemporal variables were found to be critically important, while the most relevant meteorological variables included sky cover, probability of precipitation, and maximum and minimum temperatures. Using the relevant input sets identified by the Gamma test, the developed forecasting models improve $RMSE$ for GHI by 10–15% over the reference model. $rRMSE$ range from 15% to 22% for different models constructed on 13 month data set for same-day forecasts of GHI.

In (Wang et al., 2012) an ANN model using statistical feature parameters (ANN-SFP) for short-term solar irradiance forecasting is proposed where the input vector is reconstructed with several statistical feature parameters of irradiance and ambient temperature that include (i) daily maximum value of the third order derivative of the difference between surface global irradiance and extraterrestrial global irradiance and (ii) normalized discrete difference of solar surface and extraterrestrial irradiances. These statitistical features help in reducing the model complexity because it uses five inputs. The network training is done using the Levenberg-Marquardt algorithm (LMA). Network output are 24, one for each hour of the day with the predicted irradiance value. After simulations are carried out, the proposed model is validated and compared with the conventional ANN model using historical data series (ANN-HDS) dummy model, and the results showed an inproved accuracy under variable weather conditions. The proposed model is validated using solar irradiance recorded data at Yundian Science and Technology Park grid-connected PV plant from March 2011 to December 2011.

2.4 Cloud imaginery

Using information about temporal developments of clouds may help in solar irradiance forecasting. Good performance is achieved from 30 minutes up to 6 hours using cloud motion vectors from satellite images. Cloud information from ground based sky images may be used to get irradiance forecasts with much higher spatial and temporal resolution.

Besides the deterministic factors (Earth movement relative to Sun position) clouds are the main influence factor over the changing values of solar irradiance. Clouds show a high variability in time so determining clouds position at future time is essential to forecasting accuracy. Satellite and groud-based sky images are available techniques to forecast and model solar irradiance. The basis of this method relies upon determination of clouds structure at previous time and determine future position using extrapolation. With it's high spatial and temporal resolution satellite and ground-based sky images offer the potential to derive the required information on cloud motion using motion vector fields. With this method, clouds position and structure can be determined up to 6 hours in advance. Using this information, irradiance for all sky conditions including cloudy skies maybe derived using radiative transfer models (RTM) (Diagne et al., 2013).

2.4.1 Satellite images

In (Hammer et al., 1999) satellite images are used to forecast solar surface irradiance. The satellite data provides information about cloudiness which is used to detect motion of cloud structures. Extrapolating the temporal development of the clouds, solar irradiance can be predicted from 30 minutes up to 2 hours. Images used are taken from METEOSAT satellite and an enhanced version of semi-empirical HELIOSAT method is used to derive surface irradiance. Test shown a certain improvement in forecast accuracy over persistence forecast.

A statistical fit of the relationship between a normalised parameter of the solar irradiance (such as clearness or clear sky index) and the cloud index is presented in (Zarzalejo et al., 2009). Experiment is fitted and tested using data from 28 Spanish radiometric stations (corresponding to mediterranean, semi–arid and oceanic climate areas). Local cloud index percentiles inclusion (median, first and third quartile) estimated from the whole series on each pixel improves clearly the model response and is a way to account for the local climatological characteristics of any location. The inclusion of the new variables decreases $RMSE$ value to about 17% from 21% of the expression applied in the Heliosat-2 model.

An Artificial Neural Network model is proposed in (Aguiar et al., 2015) in order to forecast GHI using ground measurement and satellite data (from Helioclim-3 database, images from Meteosat which are postprocessed to extract solar irradiation information). Forecasts horizons are intra-day, that is, from 1 hour to 6 hours ahead. Two ground stations are used to retrieve ground-based data, first one is Pozo Izquierdo a second one in Las Palmas de Gran Canaria. Best results are achieve for Pozo Izquierdo location where $RMSE$ is 15.3% for time horizon $h = 1$ and 22.6% for $h = 6$. Las Palmas $RMSE$ results are 24.3% for $h = 1$ and 36.2% for $h = 6$. The difference is due to the fact that Las Palmas has very much cloud activity in summer because of the trade winds effect, meanwhile Pozo Izquierdo, in south of the island presents a large amount of clear days.

2.4.2 Ground-based images

A method using a ground-based sky imager is presented in (Chow et al., 2011) which is used to forecast clouds development and solar irradiance. Sky cover was modeled taking sky images every 30 seconds and using a clear sky library and sunshine parameter. A two-dimensional cloud map was generated, then cloud shadows at surface were estimated. Accuracy was limited by the shortcomings of the sky imager because part of the data within each image is lost due to obscuration by the camera arm and shadowband. Also, automatic gain control of the CCD camera causes signal strength to fluctuate between images. Cloud deformation, evaporation, condensation as well as uncertainty in cloud base height causes problems in the sky conditions forecasting. Validation over four partly cloudy days showed a 70% of correct forecast for a network of six pyranometer ground stations. A 50–60% reduction in forecast error compared to persistence (cloud advection versus persistence errors) was found.

An in-house sky imager system was developed for cloud cover estimation and characterization based on a CCD camera in (Cazorla et al., 2008). The system captures a multispectral image every 5 minutes and they are analyzed using a method based on an optimized neural network classification procedure and a genetic algorithm. The genetic algorithm is used to find out the optimal input parameters set which is found to be only 3 of the initial 18 parameters, information on the red and blue channel are found to be the most important for cloud detection. Using only these 3 parameters increases speed as well as performance of the model. The method considers three types of situations: clear sky, opaque clouds and thin clouds. The images are divided into different regions and the percentage of clouds in those regions are found. The classification algorithm is validated at two levels: image level (using clouds information from the METAR register from the closest meteorological station) and pixel level (determining if the final classification is correct). Typical dust events in the south of Spain, the so called “calima” are another obstacle in the forecasting procedure.

2.5 Numerical Weather Prediction Models

Numerical weather prediction models are based on dynamical equations and can predict atmosphere conditions up to several days ahead from initial conditions. Models covering the whole Earth are the base for all others. The model equations and inputs are discretized on a three-dimensional grid that extends vertically from the Earth’s surface. Since these models are computationally very intensive, only a few are currently in operation (Traunmüller and Steinmaurer, 2010).

Models runs are tipically started two to four times per day, for example at 0, 6, 12 and 18 hours. Their initial conditions are derived from satellite, radar, ground stations, etc. and are then processed and interpolated to the 3D grid. The resolution of global NWP models are relatively coarse with grid spacings of the order of 40 to 90 kms, in order to limit computational cost. Mesoscale or limited area models are NWP models that cover a limited geographical area.

2.6 Mesoscale models

Mesoscale meteorology refers to the study of weather systems which are smaller than synoptic scale systems and larger than microscale and storm-scale cumulus systems. As said before, mesoscale models are NWP models that cover a limited geographical area but with higher resolution, to account for local terrain a weather characteristics. Initial conditions for these models can be extracted from the global models. Tipically, horizontal dimensions range from 5 kilometers to several hundreds kilometers. Example of mesoscale models are the Weather Research and Forecasting (WRF) model, Mesoscale Atmospheric Simulation System (MASS), Skiron-CENER and HIRLAM-CIEMAT (Perez et al., 2013).

HIRLAM (HIgh Resolution Limited Area Model) is a numerical short-range weather forecasting system.HIRLAM forecasts directly the following variables: ambient temperature, horizontal and vertical wind components, humidity, cloud water, pressure and geopotential height. The initial condition for the model is derived from direct observations and extrapolation of these variables. The rest of variables are derived from these prognostic variables. Since HIRLAM is a regional model, it needs a global model to act as host model and to provide lateral boundary conditions.

In (Lara-Fanego et al., 2012) the forecasts reliability of three days ahead global horizontal irradiance (GHI) and direct normal irradiance (DNI) provided by the WRF mesoscale atmospheric model for Andalusia (southern Spain) is evaluated. GHI forecasts were produced directly by the model while DNI forecasts were produced using the WRF outputs and satellite data with aerosols and ozone information. Hourly time scale and 3 km spatial resolution estimates were tested with measurements from four radiometric stations data from years 2007 and 2008. The evaluation was carried out for different forecast horizons (1, 2 and 3 days ahead), the different seasons of the year and three different sky conditions (clear, cloudy and overcast). For 24 ahead GHI forecasts, MBE was 2% for clear skies and 18% for cloudy conditions, however, the MBE for DNI increased up to 10% and 75% for same both conditions. $RMSE$ for GHI ranged from under 10% for clear skies to 50% for cloudy conditions. $RMSE$ for DNI ranged from 20% to 100% for the clear sky and cloudy conditions respectively. From these results it can be noted that DNI has a high sensitivity to the sky conditions and that $RMSE$ and MBE increase with cloudy conditions. Model was found to perform quite well with clear sky conitions but performance in cloudy conditions was similar to persistence model.

2.7 Errors of the models previously proposed

For larger forecast horizons, about 4 to 6 hours, numerical weather prediction (NWP) models tipically produce better results that satellite based mehods. Depending on the forecast horizon and available data, combined aproaches can be used to get optimized forecasts.

In (Kostylev et al., 2011) different forecasting models are analyzed for different time scales showing that satellite based images can provide accurate forecasts up to 6 hours ahead with temporal granularity of minutes producing better results than NWP-based and persistence based models. NWP has good forecasting performance at time scales from 6 hours and longer. Total sky imager allows for high spatial and temporal resolution in global horizontal irradiance (GHI) forecasts at timescales shorter than 5 minutes and a 50–60% reduction in forecast error can be achieved compared to persistence for 30 seconds ahead forecast. Several models are compared for hourly and daily forecasts and ARIMA gets the best results. Also error statistic recomendations are given, in particular, $RMSE$ should be used for time scales from seconds up to hours ahead because it better addresses the likelihood of extreme values related to ramp-up or ramp-down events, while MAE is recommended for day ahead short term forecasting as it better relates to unit commitment and energy trading by integrating absolute difference between expected and observed measures. For medium to long-term forecasts MBE is more useful for planning purposes than variability metrics because it relates the best to the assessment of total forecasted power generation capacity.

Among the works revised for this study two are of special interest because they are based on mediterranean locations as well as this syudy.

In (Perez et al., 2013), data from three cities of southern Spain are analyzed (Córdoba, Granada y Huelva), data from other locations (US, Canada and Europe) are analyzed but are not presented here. Forecast models used for these cities data are:

- European Centre for Medium-Range Weather Forecast (ECMWF) (Lorenz et al., 2009).
- WRF-UJAEN, a model of Weather Research Forecast (WRF) operated at University of Jaén (Lara-Fanego et al., 2012).
- The regional weather forecasting system Skiron operated and combined with

statistical postprocessing based on learning machines at Spain's National Renewable Energy Center (CENER).

- The High Resolution Limited Area Model (HIRLAM) operational model from the Spanish weather service AEMET combined with a statistical post-processing at CIEMAT (HIRMAL-Ciemat) (HIRLAM, 2016).

Table 2.1 shows the results for these methods, it can be seen that ECMWF outperforms the rest of the models with a composite $RMSE$ value of 22% and a best composite MAE value of 13%.

In (Pierro et al., 2015) two postprocessing approaches are tested, based on Model Output Statistics (MOS). First (MOSRH) one is a new physical based algorithm that improves the forecast of the concentration of water vapour in the atmosphere that aimes to reproduce the absortion curves in a more realistic way. The second one (MOSNN) is based on stochastic learning algorithms that use Aritificial Neural Networks ensemble to correct NWP bias error. The MOSNN is used to refine the MOSRH output, therefore the final model is a hybrid model called *MOS cascade*. The *MOS cascade* was tested with data from Rome and Lugano. Lugano has greater meteorological variability, therefore it has a greater error values. For Rome, MOSRH performs a little better ($RMSE$ value of 29%) than MOSNN (31%). For Lugano, MOSNN performs very much better than MOSRH ($RMSE$ values of 38% vs 47% respectively). Similar differences are present in MAE values, with 18% for MOSRH and 19% for MOSNN in the Rome case, and 32% for MOSRH and 27% for MOSNN in the Lugano case.

2.8 Conclusions

In this chapter several models to forecast solar global radiation proposed by different authors are presented. These models are from a wide range of approaches: statistical data mining, satellite images, ground-based images, numeric weather prediction and mesoscale models. For each model a brief description of the proposal and the obtained errors have been described.

Previously reported errors to forecast solar radiation for different Spanish locations and for some other locations with similar meteorological conditions are also presented. The root mean square errors (relative) of these proposed methods range between 20% to 47%. Hence, improving models for predicting solar radiation is a

Location	Approach	*rmse in* W/m^2	*mae in* W/m^2	*bias in* W/m^2
Córdoba	ECMWF	23%	15%	-2%
Granada	ECMWF	23%	13%	2%
Huelva	ECMWF	20%	12%	0%
Southern Spain	ECWMF	22%	13%	0%
Córdoba	CENER	26%	16%	2%
Granada	CENER	25%	16%	2%
Huelva	CENER	26%	17%	-4%
Southern Spain	CENER	25%	16%	-1%
Córdoba	HIRLAM	26%	19%	-6%
Granada	HIRLAM	32%	25%	-16%
Huelva	HIRLAM	26%	19%	-7%
Southern Spain	HIRLAM	29%	21%	-10%
Córdoba	WRF-UJAEN	28%	15%	9%
Granada	WRF-UJAEN	27%	14%	7%
Huelva	WRF-UJAEN	25%	13%	4%
Southern Spain	WRFUJAEN	26%	14%	6%
Rome	MOSRH	29%	18%	
Rome	MOSNN(MOSRH)	31%	19%	
Lugano	MOSRH	47%	32%	
Lugano	MOSNN(MOSRH)	38%	27%	

Table 2.1: Results for different models tested in southern Spain and Italy

key issue.

Bibliography

Aguiar, L. M., Pereira, B., David, M., Díaz, F., Lauret, P., 2015. Use of satellite data to improve solar radiation forecasting with bayesian artificial neural networks. Solar Energy 122, 1309 – 1324.
URL http://www.sciencedirect.com/science/article/pii/S0038092X15005927

Aguiar, R., Collares-Pereira, M., 1992. T.a.g: A time dependent autoregressive gaussian model for generating synthetic hourly radiation. Solar Energy 49(3), 167–174.

Aguiar, R., Collares-Pereira, M., Conde, J., 1988. Simple procedure for generating sequences of daily radiation values using a library of markov transition matrices. Solar Energy 4. (3), 269–279.

Bartoli, B., Coluaai, B., Cuomo, V., Francesca, M., Serio, C., 1983. Autocorrelation of daily global solar radiation. Il nuovo cimento 40, 113–122.

Box, G., Jenkins, G., 1976. Time Series Analysis forecasting and control. Prentice Hall.

Brinkworth, B., 1997. Autocorrelation and stochastic modelling of insolation sequences. Solar Energy 19, 343–347.

Brockwell, P., Davis, R. A., 2002. Introduction to Time Series and Forecasting. Springer Texts in Statistics.

Cazorla, A., Olmo, F. J., Alados-Arboledas, L., Jan 2008. Development of a sky imager for cloud cover assessment. J. Opt. Soc. Am. A 25 (1), 29–39.
URL http://josaa.osa.org/abstract.cfm?URI=josaa-25-1-29

Chang, T., 2009. Output energy of a photovoltaic module mounted on a single-axis tracking system. Applied Energy 86, 2071–2078.

Chow, C. W., Urquhart, B., Lave, M., Dominguez, A., Kleissl, J., Shields, J., Washom, B., 2011. Intra-hour forecasting with a total sky imager at the {UC} san diego solar energy testbed. Solar Energy 85 (11), 2881 – 2893.
URL http://www.sciencedirect.com/science/article/pii/S0038092X11002982

Diagne, M., David, M., Lauret, P., Boland, J., Schmutz, N., 2013. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. Renewable and Sustainable Energy Reviews 27, 65 – 76.
URL http://www.sciencedirect.com/science/article/pii/S1364032113004334

Glahn, H. R., Lowry, D. A., 1972. The use of model output statistics (mos) in objective weather forecasting. Journal of applied meteorology 11 (8), 1203–1211.

Gooijer, J. G. D., Hyndman, R. J., 2005. 25 years of iif time series forecasting: A selective review. Monash Econometrics and Business Statistics Working Papers 12/05, Monash University, Department of Econometrics and Business Statistics.
URL http://econpapers.repec.org/RePEc:msh:ebswps:2005-12

Graham, V., Hollands, K., Unny, T., 1988. A time series model for kt with application to global synthetic weather generation. Solar Energy 40, 83–92.

Hammer, A., Heinemann, D., Lorenz, E., Lückehe, B., 1999. Short-term forecasting of solar radiation: a statistical approach using satellite data. Solar Energy 67 (1–3), 139 – 150.
URL http://www.sciencedirect.com/science/article/pii/S0038092X00000384

Heck, P., Takle, E., 1987. Objective forecasts of solar radiation and temperature. Iowa State Journal of Research 62, 29–42.

HIRLAM, 2016. http://hirlam.org.

Jensenius, J., 1989. Insolation forecasting. Solar Resources, MIT Press, Cambridge, 335–349.

Jensenius, J., Cotton, G., 1981. The development and testing of automated solar energy forecasts based on the model output statistics (mos) technique. In: 1st Workshop on terrestrial solar resource forecasting and on use of satellites for terrestrial solar resource assessment, Washington, DC.

Ji, W., Chee, K. C., 2011. Prediction of hourly solar radiation using a novel hybrid model of {ARMA} and {TDNN}. Solar Energy 85 (5), 808 – 817.

URL http://www.sciencedirect.com/science/article/pii/S0038092X11000259

Koca, A., Oztop, H. F., Varol, Y., Koca, G. O., 2011. Estimation of solar radiation using artificial neural networks with different input parameters for mediterranean region of anatolia in turkey. Expert Systems with Applications 38 (7), 8756 – 8762.
URL http://www.sciencedirect.com/science/article/pii/S0957417411001059

Kostylev, V., Pavlovski, A., et al., 2011. Solar power forecasting performance–towards industry standards. In: 1st International Workshop on the Integration of Solar Power into Power Systems Aarhus, Denmark.

Lara-Fanego, V., Ruiz-Arias, J., Pozo-Vázquez, D., Santos-Alamillos, F., Tovar-Pescador, J., 2012. Evaluation of the {WRF} model solar irradiance forecasts in andalusia (southern spain). Solar Energy 86 (8), 2200 – 2217, progress in Solar Energy 3.
URL http://www.sciencedirect.com/science/article/pii/S0038092X11000582

Lorenz, E., Remund, J., Müller, S. C., Traunmüller, W., Steinmaurer, G., Pozo, D., Ruiz-Arias, J. A., Fanego, V. L., Ramirez, L., Romeo, M. G., et al., 2009. Benchmarking of different approaches to forecast solar irradiance. In: 24th European photovoltaic solar energy conference, Hamburg, Germany. Vol. 21. p. 25.

Luque, A., Hegedus, S., 2002. Handbook of photovoltaic science and engineering. John Wiley & Sons Ltd., Berlin.

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.
URL http://www.sciencedirect.com/science/article/pii/S0038092X11000193

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.
URL http://www.sciencedirect.com/science/article/pii/S0038092X10000782

Mora-López, L., de Cardona, M. S., 1998. Multiplicative arma models to generate hourly series of global irradiation. Solar Energy 63, 283–291.

Mora-Lopez, L., Sidrach-de Cardona, M., 1998. Multiplicative arma models to generate hourly series of global irradiation. Solar Energy 63 (5), 283–291.

Mora-López, L., Martínez-Marchena, I., Piliougine, M., Sidrach-de Cardona, M., 2011. Binding statistical and machine learning models for short-term forecasting of global solar radiation. In: Proceedings of the 10th international conference on Advances in intelligent data analysis X. IDA'11. Springer-Verlag, Berlin, Heidelberg, pp. 294–305.
URL `http://dl.acm.org/citation.cfm?id=2075337.2075367`

Perez, R., Lorenz, E., Pelland, S., Beauharnois, M., Knowe, G. V., Jr., K. H., Heinemann, D., Remund, J., Müller, S. C., Traunmüller, W., Steinmauer, G., Pozo, D., Ruiz-Arias, J. A., Lara-Fanego, V., Ramirez-Santigosa, L., Gaston-Romero, M., Pomares, L. M., 2013. Comparison of numerical weather prediction solar irradiance forecasts in the us, canada and europe. Solar Energy 94, 305 – 326.
URL `http://www.sciencedirect.com/science/article/pii/S0038092X13001886`

Perez, R., Moore, K., Stackhouse, P., 2007. Forecasting solar radiation preliminary evaluation of an approach based upon the national forecast database. Solar Energy 81(6), 809–812.

Pierro, M., Bucci, F., Cornaro, C., Maggioni, E., Perotto, A., Pravettoni, M., Spada, F., 2015. Model output statistics cascade to improve day ahead solar irradiance forecast. Solar Energy 117, 99 – 113.
URL `http://www.sciencedirect.com/science/article/pii/S0038092X15002212`

Reikard, G., 2009. Predicting solar radiation at high resolutions: A comparison of time series forecasts. Solar Energy 83 (3), 342 – 349.
URL `http://www.sciencedirect.com/science/article/pii/S0038092X08002107`

Sfetsos, A., Coonick, A., 2000. Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. Solar Energy 68 (2), 169 – 178.
URL `http://www.sciencedirect.com/science/article/pii/S0038092X9900064X`

Traunmüller, W., Steinmaurer, G., 2010. Solar irradiance forecasting, benchmarking of different techniques and applications of energy meteorology. In: Proceedings of the EuroSun 2010 conference.

Voyant, C., Paoli, C., Muselli, M., Nivet, M.-L., 2013. Multi-horizon solar radiation forecasting for mediterranean locations using time series models. Renewable and Sustainable Energy Reviews 28 (0), 44 – 52.
URL `http://www.sciencedirect.com/science/article/pii/S1364032113005030`

Wang, F., Mi, Z., Su, S., Zhao, H., 2012. Short-term solar irradiance forecasting model based on artificial neural network using statistical feature parameters. Energies 5 (5), 1355.
URL `http://www.mdpi.com/1996-1073/5/5/1355`

Zarzalejo, L. F., Polo, J., Martín, L., Ramírez, L., Espinar, B., 2009. A new statistical approach for deriving global solar radiation from satellite images. Solar Energy 83 (4), 480 – 484.
URL `http://www.sciencedirect.com/science/article/pii/S0038092X08002223`

Chapter 3

Materials and methods

3.1 Introduction

This chapter describes the different statistical and data mining models proposed for analyzing and forecasting solar global radiation. As mentioned in Chapter 2, new approaches based on data mining techniques have begun to be used in recent years for predicting solar radiation. The proposal of this thesis is to use both statistical and data mining techniques to these tasks.

From the Statistics, the characteristics of solar goblal for a day can be analyzed using the cumulative probability distribution function. The function is used to know how the values are distributed along a day. The analysis of how many different types of day there are can also be done using the Kolmogorov-Smirnov as it can be used to test the similarity between two samples.

From Data Mining, several clustering, classification and regression techniques will be tested for predicting solar radiation. The class prediction methods or classification methods allow to classify a sample in predetermined classes based on the data observed in the past. For example, it's possible to classify the days into several types based on the observation of diffetent meteorological parameters. Classification methods presented here include Decision Trees (DT) and Support Vector Machines for classificatoin (SVC). Regression methods will be also tested.

Finally some methods used to evaluate the error in predicting variables are presented. These methods allow to evaluate and compare the performance of different modeling techniques based on the error committed. In the prediction of continuous values the error can be quantified based on the difference between the observed values and the predicted values, however, in classification models, simply determining the success rate of the technique used to classify each sample in the correct class is possible.

3.2 Statistical Methods

In this section the statistical techniques used in this work are presented. The proposal is to review and analyze the possible utilization of the distribution function of values in a day and the usage of statistical tests to determine the similarity between two samples. Among those consulted in the bibliography, the proposal made in this work is to use the Kolmogorov-Smirnov two-sample test. This test allows the comparison of cumulative probability distribution functions of two different samples without the need to assume any distribution function on the underlying samples.

3.2.1 Cumulative probability distribution functions

The cumulative probability distribution function (CPDF) describes the probability of a random variable X with a given probability distribution to have a value less than or equal to x:

$$F_X(x) = \Pr(X \leq x) \tag{3.1}$$

The right hand side represents the probability that the random variable X takes on a value less than or equal to x. Therefore, the probability that X lies in the semi-closed interval $(a, b]$ where $a < b$ is:

$$\Pr(a < X \leq b) = F_x(b) - F_x(a) \tag{3.2}$$

Normally a capital F for a cumulative probability distribution function is used, while the lower-case f is used for probability distribution functions.

For a continuos random variable X the CPDF can be expressed as the integral of its probability density function $\int x$ as follows:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \tag{3.3}$$

3.2.2 Kolmogorov-Smirnov two sample test

The Kolmogorov-Smirnov test is based on Cumulative Probability Distribution Function (CPDF) and can be used to compare a sample with a reference probability distribution (one sample K-S test), or to compare two samples (two sample K-S test). The K-S test is a nonparametric test of the equality of continuous, one dimensional probability distributions.

The K-S statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples.

Let the cumulative probability distribution function (CPDF) of X as $F_X(\cdot)$ and the CPDF of Y as $F_Y(\cdot)$, i.e.

$$F_X(t) = \Pr(X \leq t), F_Y(t) = \Pr(Y \leq t) \tag{3.4}$$

according to Section 3.2.1.

Both $F_X(\cdot)$ and $F_Y(\cdot)$ are assumed to be continuous. Suppose that we want to test the null hypothesis

$$\mathrm{H}_0 : F_X(\cdot) = F_Y(\cdot),$$

versus the general alternative hypothesis

$$\mathrm{H}_a : F_X(\cdot) \neq F_Y(\cdot),$$

making no parametric assumption about the shape of these CPDF's.

The test can be performed using the Kolmogorov-Smirnov statistic that compares the empirical CPDF's obtained with each sample.

Specifically, if for any real number t it's defined $\hat{F}_X(t) \equiv n^{-1} \sum_{i=1}^{n} \mathbf{I}(X_i \leq t)$ and $\hat{F}_Y(t) \equiv m^{-1} \sum_{i=1}^{m} \mathbf{I}(Y_i \leq t)$, where $\mathbf{I}(A)$ is the indicator function of event A, which takes the value 1 if A is true or 0 otherwise, then the Kolmogorov-Smirnov statistic is:

$$D_{n,m} \equiv \left(\frac{nm}{n+m}\right)^{1/2} \sup_{t\in\mathbb{R}} \left|\hat{F}_X(t) - \hat{F}_Y(t)\right|.$$

The null hypothesis is rejected with significance level α if $D_{n,m} > c_\alpha$, where c_α is a critical value that only depends on α (for details, see e.g. Rohatgi and Saleh (2001)).

3.2.3 Multivariate Linear regression

Linear Regression is the simplest and most widely used of all statistical techniques, used to model the dependence of a variable on one or more explanatory variables. Regression is a method by which a functional relationship in the real world may be described by a mathematical model and then this model can be used to explore, describe or predict the relationship. In regression analysis there is usually the independent or explanatory variables and a dependant or outcome variable. The effects of one or more independent variables combine to determine the response variable.

The relationship between independent and dependant variables can be expressed as follows:

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \tag{3.5}$$

where: Y_t is the dependant variable $X_1, X_2, ..., X_p$ are the independent variables $\beta_0, \beta_1, \beta_2, ..., \beta_p$ are de influence coefficients of the independent variables over the dependent variable.

Chosing $\beta_0, \beta_1, \beta_2, ..., \beta_p$ is a minimization problem; they are selected using a cost function which expresses the difference between observed values and predicted values. For example, using a simple polinomial expression like:

$$h_\beta(x) = \beta_0 + \beta_1 x \tag{3.6}$$

The cost function would be:

$$J(\beta_0, \beta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\beta(x^{(i)} - y^{(i)})^2 \tag{3.7}$$

Then the goal is to minimize the cost function $J(\beta_0, \beta_1)$, which can be done using the gradient descent algorithm.

3.3 Data mining techniques

3.3.1 Clustering techniques

Clustering consists on finding homogeneous groups of entities (clusters) in data sets, that is, groups of samples with similar characteristics (Mirkin, 2012). The main objectives of clustering can be: *Structuring*: representing data as a set of clusters, *Description*: clusters can be described in terms of features, *Generalization*: making general statements about the data structure, *Visualization*: representing clusters structures visually.

K-means

Several partitional and hierarchical heuristic clustering methods are proposed in the classification given by (Jain et al., 1999) for clustering a set of observations. Hierarchical algorithms recursively find nested clusters but they are not suitable for large data sets as they have quadratic or higher complexity. In contrast, partitional algorithms have lower complexity as they find all the clusters simultaneously as a partition of the data without imposing hierarchical structure. Both hierarchical and partitional approaches are based on distance or dissimilarity measures.

K-means is a clustering method proposed by MacQueen, (MacQueen, 1967) and is the most widely used partitional clustering algorithm (Celebi et al., 2013). K-means aims to partition n observations into K clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. It's an unsupervised learning method, that is, only the number of desired clusters is specified and the algorithm does the job. The algorithm is based on examining one o more characteristics of the samples and group them into categories or clusters. For calculating the distance between samples (or between samples and the “mean” of each cluster) several distance metrics can be used, and the most usual are euclidean distance, manhattan distance, etc.

The euclidean distance is defined as follows:

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \tag{3.8}$$

where p and q are two points in *Euclidean space* or, in this case, vectors of characteristics.

The algorithm, based on executing cycles, is described in Algorithm 1.

```
Input  : $K$ (the number of clusters) and the Training set
         $(\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$, where $x^{(i)} \in \mathbb{R}^n$ corresponds to $F(\lambda_i))$
Randomly initialize $K$ cluster centroids $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$;
repeat
    for $i \leftarrow 1$ to $m$ do
        $c^{(i)}$ = index $j$ of the cluster centroid $\mu_j$ closest to
        $x^{(i)}, D_{x^{(i)},\mu_j} = \min\{D_{x^{(i)},\mu_k}\}, k = 1 \cdots K$ (using Eq. 3.8)
    end
    for $i \leftarrow 1$ to $K$ do
        $\mu_i$ = the average of the points assigned to cluster $i$ (this is the new
        centroid of the cluster);
    end
until Assigned indices $c^{(i)}$ do not change;
Output: Cluster for each sample (K clusters) and K centroids.
```

Algorithm 1: K-means algorithm

For the first cycle, random samples can be chosen as prototypes of each cluster or predefined samples can be used. Specific algorithm to chose initial prototypes can also be used (Celebi et al., 2013). Because of the randomness of the prototype samples chosen for the first cycle, the algorithm does not return the same result always.

The popularity of K-means can be attributed to several reasons: is simple and easy to implement, is versatile, almost every aspect of the algorithm can be configured (initialization, distance function, termination criterion), it has a linear time and storage complexity, convergence is guaranteed and it is invariant to data ordering (Celebi et al., 2013).

Disadvantages include it can only detect compact, hyperspherical clusters that are well separated (can be solved using other distance functions), when using Euclidean distance it is sensitive to noise and outlier points, it often converges to local minimum, so it is highly sensitive to initial centers.

There are different methods to initialize the centroids, the simplest are to chose the first K vectors and chose K random vectors (sensitive to ordering). Another method called Ball and Hall takes the first centroid as the mean of the total data set and then take the following ones with the restriction of a minimum distance between the candidate and the already chosen ones. Ohter methods like Al-Daoud's density-based method and Bradley and Fayyad's method are based on partitioning the data set into subsets (Celebi et al., 2013).

3.3.2 Artificial neural networks

Artificial Neural Networks (ANN) are inspired by the nervous system of animals and is a type of supervised machine learning paradigm and automatic processing. Usually there are several interconnected neurons to produce an output from a given input but system design depends on the problem to be treated. The system must be trained to set appropiate values in the propagation, activation and transfer functions, and then, the system can be used to classify samples or predict values. The main idea, like many other systems, is to have one or several inputs and one output, so the system does a process over the inputs to produce and output.

Usually, ANN are organized in three layers called input layer, hidden layer and output layer. In the input layer there is one neuron per input parameter, the input neurons are connected with neurons in the hidden layer, and these are connected with neurons in the output layer, see figure 3.1. More complex systems will have more layers of neurons.

An ANN is typically defined by three types of parameters:

- The interconnection pattern between different layers of neurons.
- The learning process for updating the weights of the interconnections.
- The activation function that converts a neuron's weighted input to its output activation.

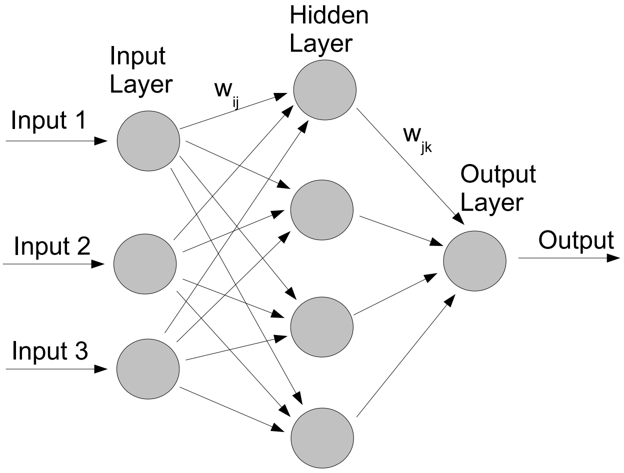


Figure 3.1: Neuronal Network

The interconnection pattern for each neuron can be described with equation 3.9, where f is the transfer function with yields an output based on the value of x_j which is transmitted through a connection that multiplies its strength by a weight $w_{i,j}$ and the resulting product is the argument for f (Mubiru, 2011). That is, the output of each neuron is a function of the inputs of that neuron. f is a transfer function; unit step (threshold), sigmoid, piecewise linear, and gaussian are the most common transfer functions.

$$y_i = f(\sum_{j=1}^{n} x_j w_{i,j}) \tag{3.9}$$

where $w_{i,j}$ is the weight of the connection between neuron j of previous layer and neuron i in the current layer.

An ANN is composed of several interconnected neurons, each one has several inputs and one output, that output depends on tree factors:

- Propagation function, typically, the sum of each input multiplied by a specific weight.
- Activation function, can modify the propagation function output and is op-

tional.

- Transfer function, that modifies the activation function, typically to fit output in a range.

ANN are a powerful tool to solve many problems, here are some of the advantages of ANN:

- Learning, ANN can learn in the learning stage in a supervised learning way.
- Self organizing, ANN creates his own representation of the problem, user don't have to worry about that.
- Fault tolerance, information is stored in redundant mode, so it will still function if is partially damaged.
- Flexibility, ANN can deal with minor changes in the inputs.
- Real Time, due to it's parallel structure, it can be implemented to be very fast in response.

Learning is the most interesant possibility in neural networks. Given a specific task to solve, and a class of functions F, learning means using a set of observations to find $f^* \in F$ which solves the task in some optimal sense.

A cost function must be defined, $C : F \to \Re$, and for the optimal solution, $f^*, C(f^*) \leq C(f) \forall f^* \in F$, i.e., no solution has a cost less than the cost of the optimal solution. The cost function determines how far away a particular solution is from an optimal solution to the problem to be solved. Learning algorithms search through the solution space to find a function that has the smallest possible cost. For applications where the solution is dependent on some data, the cost must necessarily be a function of the observations, otherwise it would not be modeling anything related to the data.

While it is possible to define some arbitrary cost function, frequently a particular cost will be used, either because it has desirable properties (such as convexity) or because it arises naturally from a particular formulation of the problem.

There are three learning paradigms, supervised learning, unsupervised learning and reinforcement learning.

- Supervised learning. A set of example pairs is given, $(x, y), x \in X, y \in Y$ and the objetive is to find a function $f : X \to Y$ in the allowed class of functions that matches the examples. A commonly used cost is the mean-squared error, which tries to minimize the average squared error between the network's output, $f(x)$ and the target value y over all the example pairs. When gradient descent is used to minimize the cost for the class of neural networks called multilayer perceptrons, the well-known back-propagation algoritm is being used.

- Unsupervised learning. The given cost function can be any function of the given data x and the network's output f. The cost function depends on what is being modeled and the basic assumptions.

- Reinforcement learning. Usually, the data is not given, but, is generated by the agent's interaction with the environment. At each point in time t, the agent performs an action y_t and the environment generates an observation x_t and an instantaneous cost c_t, according to some (usually unknown) dynamics. The aim is to discover a policy for selecting actions that minimizes some measure of a long-term cost; i.e., the expected cumulative cost.

Training a neural network model means selecting one model from the set of allowed models that minimizes the cost criterion. Most of the models can be viewed as a straightforward application of optimization theory and statistical estimation. Most of the algorithms used in training artificial neural networks employ some form of gradient descent, taking the derivative of the cost function with respect to the network parameters and then changing those parameters in a gradient-related direction.

The training data set consists of N training patterns (x_p, t_p), where p is the pattern number. x_p is the input vector with dimension N and t_p is the desired output vector with dimension M. y_p is the network output vector for the pth pattern. The Levenberg-Marquardt (LM) method can be used for training the ANN and is one of the most popular algorithms. LM updates the neurons weight as follows:

$$\Delta w = -\left[\mu I + \sum_{p=1}^{P} J^p(w)^T J^p(w)\right]^{-1} \nabla E(w) \tag{3.10}$$

where $J^p(w)$ is the Jacobian matrix of the error vector $E^p(w)$ evaluated in w, and I is the identity matrix. The vector error $E^p(w)$ is the error of the network for pattern p: $E^p(w) = T^p - O^p(w)$.

Some commonly used methods for training neural networks include evolutionary methods, gene expression programming, simulated annealing, expectation maximization, non-parametric methods and particle swarm optimization.

3.3.3 Support Vector Machines

Support Vector Machines (SVM) is a supervised learning model (introduced by Vapnik et al. in 1992 (Berthold and Hand, 2003)), aimed to be an efficient way of learning 'good' separating hyperplanes in a high dimensional feature space, capable of analyzing and recognizing patterns, and can be used for classification and non-linear regression analysis. A SVM is a linear machine based on Structural Risk minimization (SRM) method and on statistical learning theory and can provide good performance in pattern recognition problems without problem domain knowledge (Gorunescu, 2011). In the simplest case, given a set of training samples, each belonging to one of two categories, an SVM builds a model that assigns new samples to one of the two categories.

The fundamental concept of a SVM, based on Cover's Theorem (Cover, 1965), is that, while samples in the input (low-dimensional) space could not be linearly separable, mapping them into a sufficiently high-dimensional space (feature space) using a nonlinear mapping function $\varphi(x)$ could more probably be linearly separable.

A SVM is based on the implementation of the following two steps:

- Mapping the training points by a nonlinear function $\varphi(x)$ to a high-dimensional space in which the training points are linearly separable.
- Determining the optimal separation hyperplane that maximizes margin (the distance between points of each category).

In this work, SVM implementation from Chih-Chung Chang and Chih-Jen Lin is used (Chang and Lin, 2011), which is available at http://www.csie.ntu.edu.tw/ cjlin/libsvm/.

This implementation supports the following types of tasks:

- SVC: Support Vector Classification, two-class and multi-class.
- SVR: Support Vector Regression.
- One-class SVM

Four types of kernels are supported by LibSVM: linear, polynomial, radial basis function and sigmoid function. Using LIBSVM requires two steps: first, training a data set to construct a model and second, using the model to predict information from a testing data set. Getting output probabily estimates is also posible with this library.

In the following section details abour ν-SVC and ν-SVR are provided as these are the variants of LibSVM that are used in this thesis.

ν-Support Vector Classification

Given training vectors $x_i \in R^n, i = 1..., l$ in two clases, and given $y \in R^l$, such that $y_i \in 1, -1$, the primal optimization problem is:

$$\begin{aligned} \min_{\omega,b,\xi,\rho} \quad & \frac{1}{2}\omega^T\omega - \nu\rho + \frac{1}{l}\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & y_i(\omega^T\phi(x_i) + b) \geq \rho - \xi_i, \\ & \xi_i \geq 0, i = 1, \ldots, l,\ \rho \geq 0. \end{aligned} \tag{3.11}$$

The dual problem is

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\alpha^T Q\alpha \\ \text{subject to} \quad & 0 \leq \alpha_i \leq 1/l, \quad i = 1, \ldots, l, \\ & e^T\alpha \geq \nu,\ y^T\alpha = 0, \end{aligned} \tag{3.12}$$

where $Q_{i,j} = y_i y_j K(x_i, x_j)$. (Chang and b. Lin, 2001) showed that problem 3.12 is feasible if and only if

$$\nu \leq \frac{2min(\#y_i = +1, \#y_i = -1)}{l} \leq 1, \tag{3.13}$$

so the range for ν is smaller than (0,1].

ν-Support Vector Regression

For $\nu - SVR$ also a parameter $\nu \in (0,1]$ is used, like in $\nu - SVC$, to control the number of support vectors. Also, the parameter $\epsilon - SVR$ becomes a parameter here. The problem to solve is:

$$\begin{aligned} \min_{\omega,b,\xi,\xi^*,\epsilon} \quad & \frac{1}{2}\omega^T\omega + C(\nu\epsilon + \frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*)) \\ \text{subject to} \quad & (\omega^T\phi(x_i) + b) - z_i \leq \epsilon + \xi_i, \\ & z_i - (\omega^T\phi(x_i) + b) \leq \epsilon + \xi_i^*, \\ & \xi_i, \xi_i^* \geq 0, i = 1, \dots, l,\ \epsilon \geq 0. \end{aligned} \tag{3.14}$$

The dual problem is:

$$\begin{aligned} \min_{\alpha,\alpha^*} \quad & \frac{1}{2}(\alpha - \alpha^*)^T Q(\alpha - \alpha^*) + z^T(\alpha - \alpha^*) \\ \text{subject to} \quad & e^T(\alpha - \alpha^*) = 0, e^T(\alpha + \alpha^*) \leq C\nu, \\ & 0 \leq \alpha_i, \alpha_i^* \leq C/l,\ i = 1, \dots, l. \end{aligned} \tag{3.15}$$

The aproximate function is

$$\sum_{i=1}^{l}(-\alpha_i + \alpha_i^*)K(x_i, x) + b \tag{3.16}$$

As in $\nu - SVC$, the inequality $e^T(\alpha + \alpha*) \leq C\nu$ can be replaced by an equailty. Because users usually choose a small value for C, C/l may be too small. Then, in LibSVM, the users specifies $\overline{C} = C/l$, and LibSVM solves the following problem:

$$\begin{aligned} \min_{\alpha,\alpha^*} \quad & \frac{1}{2}(\alpha-\alpha^*)^T Q(\alpha-\alpha*) + z^T(\alpha-\alpha^*) \\ \text{subject to} \quad & e^T(\alpha-\alpha^*) = 0 \ , e^T(\alpha+\alpha^*) = \overline{C}l\nu, \\ & 0 \leq \alpha_i, \alpha_i^* \leq \overline{C}, \ i = 1, \dots, l. \end{aligned} \tag{3.17}$$

In (Chang and b. Lin, 2002) was proved that $\epsilon - SVR$ with parameters $(\overline{C}, \epsilon)$ has the same solution as $\nu - SVR$ with parameters $(l\overline{C}, \nu)$.

3.3.4 Decision Trees

According to (Rokach and Maimon, 2008), in machine learning environment prediction methods are commonly referred to as supervised learning, which stands opposed to unsupervised learning (which refers to modeling the distribution of instances in a input space). Supervised methods are those that attempt to establish a relationship between input attributes (also called independent variables) and a target attribute (also called dependen variable). The discovered relationship is represented in a structure called Model. The Model can be used to predict values of the dependent variable knowing the values of independent variables. Supervised models can be divided into two subclasses: Classification Models and Regression Models. Regression Models can predict a real (continuous) value while Classification Models can map the input values into predefined classes.

A decision tree structure can be described as a labeled directed acyclic graph where all nodes except the root have a single parent and can have zero, one or more children. Nodes with no children are called leaf nodes (also known as “terminal” or “decision” nodes). In Figure 3.2 a example of decision tree is shown. A decision tree is a predictive model that uses a tree like structure to predict values from observations. It is useful in the areas of statistics, data mining and machine learning. The predicted values can be a finite set of values (classification trees) or can be continuous values (regression trees).

In a decision tree structure each internal node denotes a test of an attribute, each branch represents an outcome of the test and each leaf node holds a class label and the topmost node in a tree is the root node (Han and Kamber, 2000). In the simplest case, each test considers a single attribute and the instance space is partitioned according to the attribute values. In the case of numeric attributes, the conditions refers to a range.

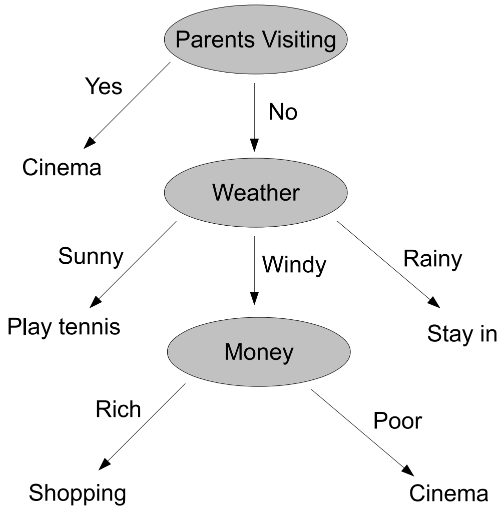


Figure 3.2: Decision Tree

Classification trees can be used to clasify objects into a predefined set of predefined classes based on their attributes. Classification trees are frequently used in applied fields such as finance, marketing, engineering and medicine.

Given a tuple X, for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree, tracing a path from the root to a leaf node which holds the class predicted for that tuple. Decision tree construction does not require any domain knowledge, can handle multidimensional data and the representation of acquired knowledge in tree form is intuitive and easy for humans to assimilate. The learning process is simple and fast, and, in general, the accuracy is good, however, success depends on the data used.

Tree size or complexity can have a important effect on its accuracy. Normally ,tree complexity is measured by one af these metrics: total number of nodes, total number of leaves, tree depth and number of attrinutes used. Attributes are tipically one of two types: nominal (values of an unordered set) and numeric (real numbers or so). The domain of an attribute can be denoted as $dom(a_i) = \left\{v_{i,1}, v_{i,2}, ...v_{i,|dom(a_i)|}\right\}$ where $|dom(a_i)|$ is the attribute cardinality. For the target attribute the domain can be represented as $dom(y) = \left\{c_1, c_2, ...c_{|dom(y)|}\right\}$. The instance space can be defined as the Cartesian product of all the input attributes domains: $X = dom(a_1) \times dom(a_2) \times ... \times dom(a_n)$; and the universal instance space also includes the target attribute: $U = X \times dom(y)$.

The training set is a set of unordered tuples, which can be represented as $S(B) = (\langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle, ... \langle x_m, y_m \rangle)$ where $x_q \in X$ and $y_q \in dom(y)$. An induction algorithm (or inducer) is an entity capable of construct a model (classifier) that generalizes the relationship between the input attributes and the target attribute. Using this classifier it is possible to prediuct the target value of a tuple x_q.

The classifier can be used to classify an unseen tuple either by explicitly assigning it to a certain class or by providing a vector of probabilities representing the conditional probability of the given instance to belong to each class. Induction of an optimal decision tree from a given data set is considered to be a difficult task (NP-hard o NP-complete, Rokach and Maimon (2008)), so it is only feasible in small problems. Heuristic methods are necessary in order to construct a optimal decision tree. Methods for this can be classified into two categories, top-down and bottom-up. Examples of top-down (the most frequent) are ID3, C4.5 and CART.

Split criterion is used to choose which attribute is used to create child nodes from a given node in a decision tree. Next, split cliterion is explained. Given a set of experiencies E we define M as a measure of disorder in E. If we classify by an attribute A then a child nodes will be produced, with partial measures M_j $(j = 1, \ldots, a)$. With these we calculate the new measure in the parent node, using the ponderate measure:

$$M(A) = \sum_{j=1}^{a} prob(A_j M_j \tag{3.18}$$

where

$$prob(A_j) = \frac{\text{number of experiences for } j \text{ value for attribute } A \text{ at this node}}{\text{number of experiences at this node}} \tag{3.19}$$

When classifying with attribute A, the disorder decrement would be:

$$\Delta(A) = M - M(A) \tag{3.20}$$

The attribute with the bigger decrement will be the choosen one by the Split Criterion.

Next, the complete algorithm for constructing a decision tree is presented.

```
TreeGrowing($S, A, y, SplitCriterion, StoppingCriterion$)
Where:
$S$ - Training Set
$A$ - Input Feature Set
$y$ - Target Feature
$SplitCriterion$ - the method for evaluating a certain split
$StoppingCriterion$ - the criteria to stop the growing process

Create a single tree $T$ with a single root node
IF $StoppingCriterion(S)$ THEN
   Mark $T$ as a leaf with the most common value of $y$ in $S$ as label
ELSE
   $\forall\ a_i \in A$ find $a$ that obtain the best $SplitCriterion(a_i, S)$
   Label $t$ with $a$
   FOR EACH outcome $v_i$ of $a$:
      SET $Subtree_i = TreeGrowing(\sigma_{a=v_i} S, A, y)$
      Connect the root node of $t_T$ to $Subtree_i$ with an edge that is labelled as $v_i$
   END FOR
END IF
RETURN TreePrining($S, T, y$)

TreePruning($S, T, y$)
Where:
$S$ - Training Set
$y$ - Target Feature
$T$ - The tree to be pruned

DO
   Select a node $t$ in $T$ such that pruning it maximally improve some evaluation criteria
   IF t $\neq \oslash$ THEN $T = prunet(T, t)$
UNTIL $t = \oslash$
RETURN $T$
```

Algorithm 2: Decision Tree Growing Algorithm

3.4 Metrics

Model performance is usually evaluated against a single dataset and the strategy is to select the model with the best performance for this particular dataset. An individual error is the difference between a predicted value and the corresponding observed or true value.

Next, statistical methods used to measure forecasting errors are introduced. Each individual statistic used to measure the error has advantages, also disadvantages over the others. Several authors have proposed a set of metrics to evaluate general forecasting methods evaluation and error metrics, see for example (Chen and Yang, 2004). In (Voyant et al., 2015) 20 statistical parameters to estimate the short term predictability of the global horizontal irradiation time series are reviewed. The mean absolute log-return (which the author claims to have never been used before in global radiation forecasting) proves to be very efficient.

In (B.Viorel, 2008) MBE (Mean Bias Error), $RMSE$ (Root Mean Square Error) and $MABE$ (Mean Absolute Bias Error) are cited as basic error metrics. MBE measures the systematic errors (or bias) while $RMSE$ is mostly a measure of random errors and $MABE$ is less frequently used than the two other statistics.

In this thesis several statistics are used so that the joint vision allows provide better assess of the accuracy of prediction models.

First step for calculating forecasting errors is to define the residue series, that is, the difference between observed and forecasted values:

$$e_t = \hat{Z}_t - Z_t \tag{3.21}$$

where e_t are the residue, $\hat{Z}_t$ are the forecasted values of the interest variable and Z_t are the actual values of the forecasted variable.

3.4.1 MAPE

One of the most used method is the *Mean Absolute Percentage Errror (MAPE)*, introduced in equation 3.22. To avoid cancelation from distinct signed values, absolute value is used. Multiplying by 100, this method's result is expressed as percentage:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} |\frac{e_t}{Z_t}|(x100) \tag{3.22}$$

The Advantage of this method is that it allows to directly compare error between different systems because is normalized. The disavantage is that it does not discriminate error from low values, for example, low solar radiation energy values that occur early in the morning or at the sunset when energy is low.

3.4.2 MSE

The *Mean Square Error (MSE)*, also used to evaluate the output of data mining models, is estimated according to equation 3.23:

$$MSE = \frac{\sum_{t=1}^{m}(X_t - \hat{X}_t)^2}{m}. \tag{3.23}$$

3.4.3 RMSE

The next method introduced here is *Root Mean Square Error (RMSE)*, which is described in equiation 3.24. It provides a good measure of the model's precission because it is a cuadratic magnitude, that is, gives more importance to bigger error values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2} \tag{3.24}$$

The main disadvantage of this method is that is expressed in same units as the forecasted variable so is dependent of it's dynamic margin. This method is suitable when forecasting solar radiation only because it's dynamic margin is similar in all places. Also, this method is a uncertainty measure widely used in energy and PV performance studies, so it allows comparision with other studies.

3.4.4 MAE

The *Mean Absolute Error (MAE)* is defined as sum of the absolute values of the differences between the predicted values and observed values normalized to the total number of observed values, according to the equation 3.25:

$$MAE = \frac{\sum_{t=1}^{m} |X_t - \hat{X}_t|}{m}. \tag{3.25}$$

As the name suggests, the mean absolute error is an average of the absolute error of each predicted and observed pair of values.

3.4.5 RMAE

The *Relative Mean Absolute Error (rMAE)* is defined as the sum of the absolute values of the differences between each predicted value and the corresponding observed value normalized to the sum of observed values, according to the equation 3.26:

$$rMAE = \frac{\sum_{i=1}^{m} |Z_t - \hat{Z}_t|}{\sum_{i=1}^{m} Z_t} 100(\%). \tag{3.26}$$

The rMAE allows us to know the error in the estimation of total solar radiation; it was calculated as this information is very useful for solar plants connected to the grid. The data used is the difference between the values of total predicted radiation (directly used to predict the energy produced) and the total actual radiation received, due to the penalization applied to this difference.

3.5 Available meteorological data

To carry out the entire radiation forecasting process it is necessary to have a wide range of data available and also these data must be consistent and accurate.

These data can be obtained from several sources: directly from meteorological stations, meteorological web services, etc.

Meteorological stations have the advantage that data are available quickly because it can be downloaded from the station as soon as it is recorded. On the other hand a preprocessing job must be done to correct measure errors if possible or discard these values.

Data obtained from meteorological web services are normally already preprocessed so they are error free and can be used directly or with little processing.

Next, the meteorological data sources used in this study are described.

3.5.1 Meteorological station of UMA

A meteorological station able to record several parameters is available at the Photovoltaic Systems Laboratory of the Universidad de Málaga. This station is located over the roof of the ETSI Informatica of Universidad de Málaga (latitude is 36^{o}42'54" N, Longitude 4^{o}28'39 W, 45 meters elevation).

The data acquisition system installed on the ETSI Informatica is from the manufacturer *National Instruments*, which has several platforms for process control and data capture, the one used in this work is named Compact FieldPoint.

The system is composed of several elements:

- Solar Tracker: a two axis solar tracker for direct beam solar radiation measurement (figure 3.3).
- Pyranometer (figure 3.4).
- Temperature and humidity sensor (figure 3.5).
- System CPU (figure 3.6).

These systems are modular, therefore a controller module is necessary, which contains basically the system CPU and communication interfaces: RS232, RS485, Ethernet and CompactFlash card reader. In this case the cFP-2120 module is

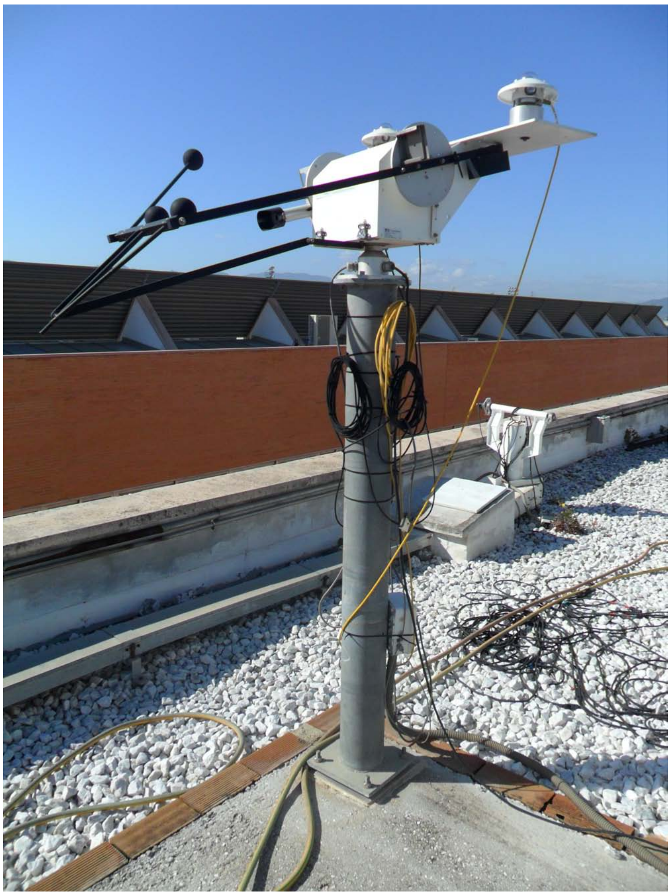


Figure 3.3: Solar tracker

installed. This module plugs into the chassis main slot which allows connection to other modules (up to 8 expansion modules are allowed plus the main controller).

Each expansion module is specialized in a particular task: relays control, current reading, voltage reading, thermometers reading, pulse counter module, current or voltage signal generation, etc. In addition to the module itself, which contains the electronics necessary to perform each function (the A/D converter that apply) next to each expansion module there is a terminal strip, which simply has the terminals for connecting wire pairs that carry signals from each one of the sensors to the FieldPoint system.

In this system there are 8 expansion modules:

- 3 cFP-RTD-124 units, for reading up to 8 (temperature) Pt100 with 4 wires

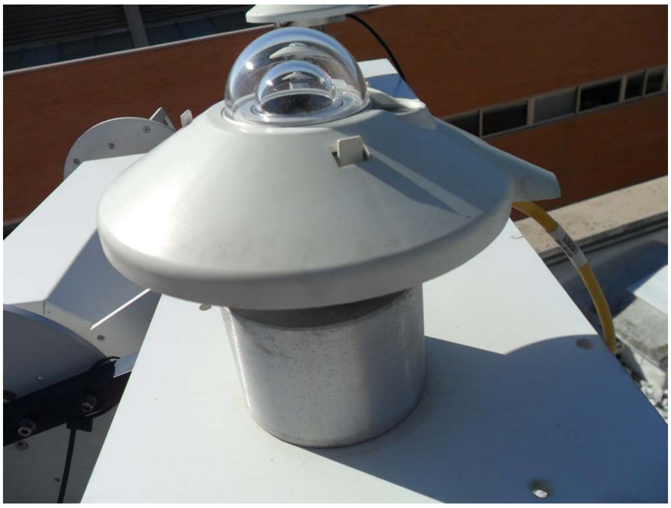


Figure 3.4: Pyranometer

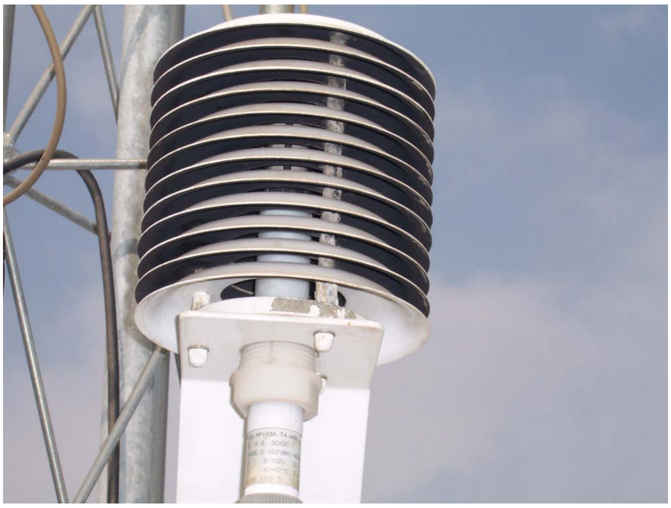


Figure 3.5: Sensor

each.

- 1 cFP-AI-111 unit, for reading up to 16 current signals (0-20 mA, 4-20 mA, -20 a 20 mA)
- 1 cFP-AI-112 unit, for reading up to 16 voltage signals (ranges from 60mV to 10V)
- 1 cFP-CTR-502 unit, for pulse counting.
- 2 cFP-RLY-425 units, with 8 relays each one that can be used as switches.

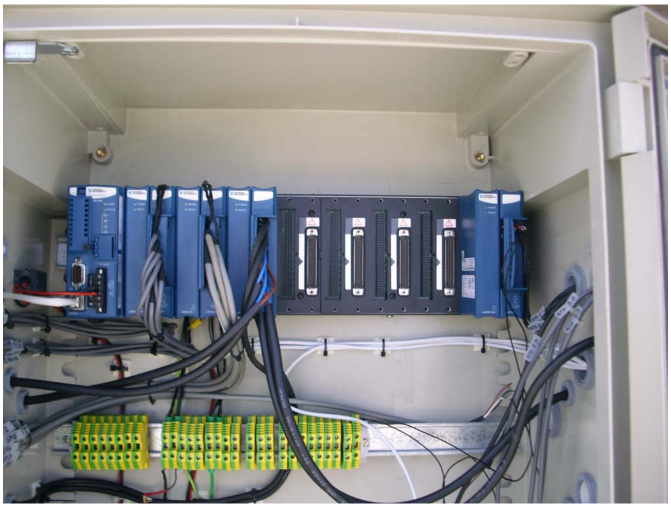


Figure 3.6: Fieldpoint CPU

The main controller cFP-2120 has the following factory default software installed:

- Operating System: Phar Lab (from IntervalZero)
- Web server
- FTP server

Datalogger functionality is not available by default, but it can be enabled using a free software from *National Instruments*, developed over LabView, which can store all measured values from modules into regular text files. These files are recored in the system's internal memory or in the memory card (*Compact Flash*). Also, this software enables remote configuration of the system using a web interface.

Recorded parameters include air temperature, humidity, atmospheric pressure and solar global horizontal radiation. The system provides data in comma separated values (CSV) format which are easy to read and process. The available data were recorded from October 2010 to December 2013, every minute, 60 records per hour, 1440 records per day in total.

Data are preprocessed to remove undesired values, such as night hour values or out of range values, measurements error, etc. The data are summarized as hourly records, so hourly means are calculated for air temperature, humidity, pressure and solar global radiation. For convenience, extraterrestrial solar horizontal radiation and clearness index are calculated and stored within the whole data set in the pre-processing.

The following estimated variables were used in the analysis and forecast processes:

- The values of clearness index for the 8 centered hours around midday were estimated.
- Meteorological data were summarized into 3-hour registers and then into one-day registers.
- The value of daily clearness index K_d

Data were summarized into 3-hour registers because predictions that can be found in public weather services are made in a 3-hour basis or (similar periods) and they can therefore be used in solar radiation forecasting models.

3.5.2 OpenWeatherMap

OpenWeatherMap is a project inspired by OpenStreetMap and Wikipedia in the sense of making information free and available for everybody. The service nowadays delivers 1 billion forecast per day. It provides open weather data for more than 200.000 cities via website and API. A wide range of data is supplied including current weather, forecasts, historical data, precipitation, wind, clouds, etc, (OpenWeatherMap, 2016).

OpenWeatherMap offers meteorological information through an API. Currently, there are eight API's for different purposes:

- Current weather data: data from over 200.000 cities available in JSON, XML or HTML format, frequently updated.
- 5 day / 3 hour forecast: for any location, forecast data for 5 days ahead, every 3 hours, available in JSON, XML or HTML format.

- 16 day / daily forecast: for any location, forecast data for 16 days ahead, daily basis, available in JSON, XML or HTML format.
- Historical data: historical weather data for more than 20.000 cities.
- UV index: current UV index and historical data for any geo location.
- Weather map layers: maps with precipitation, clouds, pressure, temperature and wind info.
- Weather stations: recent data from more than 40.000 stations around the world, search weather stations close to a geographic location.
- Bulk downloading: bulk files with current weather and forecasts from more than 20.000 cities.

Figure 3.7 shows the main page of the OpenWeatherMap web site. Figure 3.8 shows the OpemWeatherMap platform conception, structured in three layers, with de data sources in the lower layer and the providing data services in the upper layer.

Data are collected from thousands of sensors distributed all around the planet, including data from satellites and radars. Archives content data from MODIS, Landsat 7 and 8. Ground sensor network covers more than 40.000 weather stations.

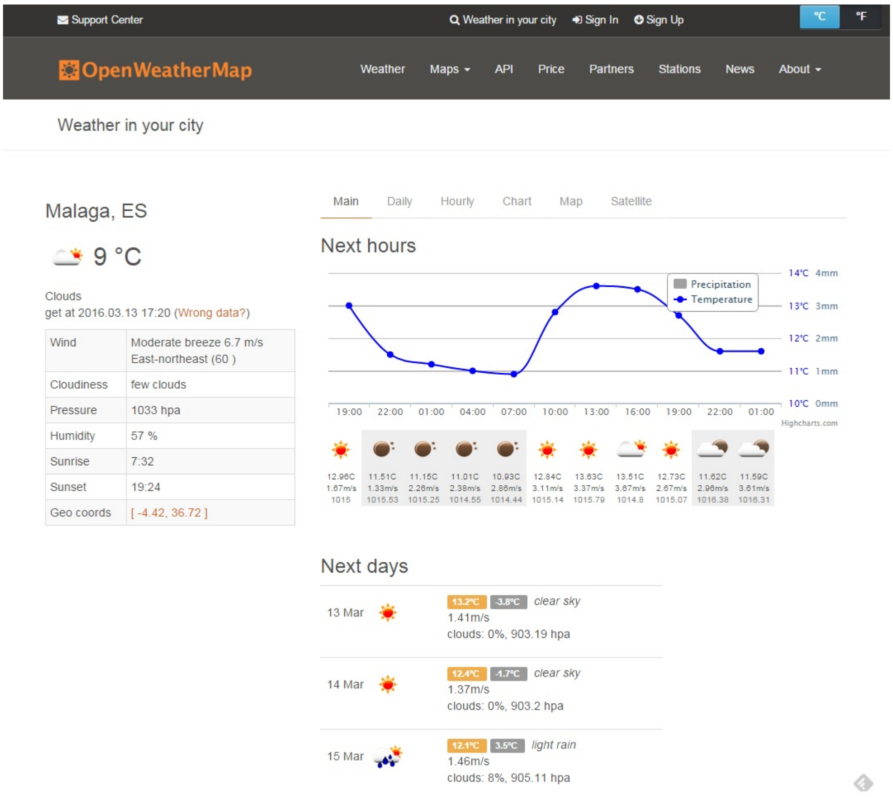
High-end technologies are used to automatically process billions of data points every second. Big data and cloud technologies enables OpenWeatherMap to work with immense data and process them right away providing the user with images, maps and data.

For the purposes of this work, current weather data and 5 day / 3 hour forecast data are downloaded and stored in a data base.

Current weather data

Current weather data can be obtained using an API directly on the OpenWeatherMap website (*www.openweathermap.org*). The url for the API call look like this:

```
http://api.openweathermap.org/data/2.5/weather?id=2514256&
    ↪ appid=b1b15e88fa797225412429c1c50c122a
```



Weather in your city

Malaga, ES

9 °C

Clouds
get at 2016.03.13 17:20 (Wrong data?)

Wind	Moderate breeze 6.7 m/s East-northeast (60)
Cloudiness	few clouds
Pressure	1033 hpa
Humidity	57 %
Sunrise	7:32
Sunset	19:24
Geo coords	[-4.42, 36.72]

Next hours

12.96C	11.51C	11.15C	11.01C	10.93C	12.84C	13.63C	13.51C	12.73C	11.62C	11.59C
1.67m/s	1.33m/s	2.26m/s	2.38m/s	2.86m/s	3.11m/s	3.37m/s	3.67m/s	2.67m/s	2.96m/s	3.61m/s
1015	1015.53	1015.25	1014.55	1014.44	1015.14	1015.79	1014.8	1015.07	1016.38	1016.31

Next days

13 Mar 13.2°C -3.8°C *clear sky*
1.41m/s
clouds: 0%, 903.19 hpa

14 Mar 12.4°C -1.7°C *clear sky*
1.37m/s
clouds: 0%, 903.2 hpa

15 Mar 12.1°C 3.5°C *light rain*
1.46m/s
clouds: 8%, 905.11 hpa

Figure 3.7: OpenWeatherMap web site

The url includes two key parameters, and API KEY for user identification purposes (example: b1b15e88fa797225412429c1c50c122a) and a station id which designates the station we want the data from, in this case 2514256.

The response is in JSON format and looks like this:

```
{"coord":{"lon":-4.42,"lat":36.72},"weather":[{"id":800,"main":"Clear","description":"clear
    ↪ sky","icon":"01n"}],"base":"cmc stations","main":{"temp":284.453,"pressure":1018.24,"
    ↪ humidity":100,"temp_min":284.453,"temp_max":284.453,"sea_level":1034.66,"grnd_level
    ↪ ":1018.24},"wind":{"speed":2.86,"deg":116.505},"clouds":{"all":0},"dt":1457829531,"sys
    ↪ ":{"message":0.0028,"country":"ES","sunrise":1457850661,"sunset":1457893413},"id
    ↪ ":2514256,"name":"Malaga","cod":200}
```

This responde includes the following parameters: temperature, pressure, humidity, wind speed and direction and the name of the station, in this case *Málaga*.

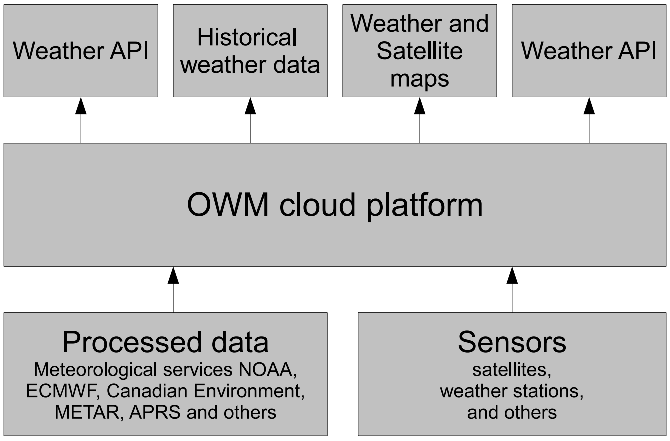

Weather API
Historical weather data
Weather and Satellite maps
Weather API
OWM cloud platform
Processed data
Meteorological services NOAA, ECMWF, Canadian Environment, METAR, APRS and others
Sensors
satellites, weather stations, and others

Figure 3.8: OpenWeatherMap platform diagram

5 day / 3 hour forecast

Similar to current weather, 5 day forecasts can be downloaded using an API on the web site. The url look like this:

```
http://api.openweathermap.org/data/2.5/forecast?id=2514256&
    ↪ appid=b1b15e88fa797225412429c1c50c122a
```

Again, the url includes the two key parameters, the API KEY of the user and the station code. The response is again in JSON format:

```
{"city":{"id":2514256,"name":"Malaga","coord":{"lon":-4.42034,"lat":36.720161},"country":"ES
    ↪ ","population":0,"sys":{"population":0}},"cod":"200","message":0.0026,"cnt":40,"list
    ↪ ":[{"dt":1457838000,"main":{"temp":284.45,"temp_min":284.45,"temp_max":284.452,"
    ↪ pressure":1017.3,"sea_level":1033.72,"grnd_level":1017.3,"humidity":100,"temp_kf":0},"
    ↪ weather":[{"id":800,"main":"Clear","description":"clear sky","icon":"01n"}],"clouds":{"
    ↪ all":0},"wind":{"speed":3.31,"deg":85.0045},"sys":{"pod":"n"},"dt_txt":"2016-03-13
    ↪ 03:00:00"},{"dt":1457848800,"main":{"temp":283.94,"temp_min":283.94,"temp_max
    ↪ ":283.944,"pressure":1017.34,"sea_level":1033.8,"grnd_level":1017.34,"humidity":100,"
    ↪ temp_kf":0},"weather":[{"id":800,"main":"Clear","description":"clear sky","icon":"01n
    ↪ "}],"clouds":{"all":0},"wind":{"speed":3.86,"deg":65.5037},"sys":{"pod":"n"},"dt_txt
    ↪ ":"2016-03-13 06:00:00"},{"dt":1457859600,"main":{"temp":285.67,"temp_min":285.67,"
    ↪ temp_max":285.671,"pressure":1018.28,"sea_level":1034.62,"grnd_level":1018.28,"humidity
    ↪ ":100,"temp_kf":0},"weather":[{"id":800,"main":"Clear","description":"clear sky","icon
    ↪ ":"01d"}],"clouds":{"all":0},"wind":{"speed":4.06,"deg":63.5053},"sys":{"pod":"d"},"
    ↪ dt_txt":"2016-03-13 09:00:00"},{"dt":1457870400,"main":{"temp":287.15,"temp_min
    ↪ ":287.15,"temp_max":287.151,"pressure":1018.11,"sea_level":1034.23,"grnd_level
    ↪ ":1018.11,"humidity":94,"temp_kf":0},"weather":[{"id":800,"main":"Clear","description
    ↪ ":"clear sky","icon":"01d"}],"clouds":{"all":0},"wind":{"speed":2.62,"deg":120.515},"
    ↪ sys":{"pod":"d"},"dt_txt":"2016-03-13 12:00:00"},{"dt":1457881200,"main":{"temp
    ↪ ":286.97,"temp_min":286.967,"temp_max":286.97,"pressure":1016.38,"sea_level":1032.33,"
    ↪ grnd_level":1016.38,"humidity":95,"temp_kf":0},"weather":[{"id":800,"main":"Clear","
    ↪ description":"clear sky","icon":"01d"}],"clouds":{"all":0},"wind":{"speed":2.41,"deg
    ↪ ":149.001},"sys":{"pod":"d"},"dt_txt":"2016-03-13 15:00:00"},{"dt":1457892000,"main":{"
    ↪ temp":286.12,"temp_min":286.12,"temp_max":286.124,"pressure":1015.41,"sea_level
    ↪ ":1031.44,"grnd_level":1015.41,"humidity":100,"temp_kf":0},"weather":[{"id":800,"main
    ↪ ":"Clear","description":"clear sky","icon":"01d"}],"clouds":{"all":0},"wind":{"speed
    ↪ ":1.06,"deg":157},"sys":{"pod":"d"},"dt_txt":"2016-03-13 18:00:00"},{"dt":1457902800,"
    ↪ main":{"temp":284.57,"temp_min":284.57,"temp_max":284.572,"pressure":1015.64,"sea_level
    ↪ ":1031.9,"grnd_level":1015.64,"humidity":100,"temp_kf":0},"weather":[{"id":800,"main":"
    ↪ Clear","description":"clear sky","icon":"01n"}],"clouds":{"all":0},"wind":{"speed
    ↪ ":0.92,"deg":146.004},"sys":{"pod":"n"},"dt_txt":"2016-03-13 21:00:00"},{"dt
    ↪ ":1457913600,"main":{"temp":284.02,"temp_min":284.02,"temp_max":284.021,"pressure
    ↪ ":1015.46,"sea_level":1031.74,"grnd_level":1015.46,"humidity":100,"temp_kf":0},
\vdots
"weather":[{"id":801,"main":"Clouds","description":"few clouds","icon":"02n"}],"clouds":{"all
    ↪ ":24},"wind":{"speed":4.71,"deg":91.0017},"sys":{"pod":"n"},"dt_txt":"2016-03-18
    ↪ 00:00:00"}]}
```

OpenWeatherMap database

A simple dababase is used to store OpenWeatherMap's current and forecasted values. An application has been developed to automatically read data from web site and and store it into the database.

3.5.3 Spanish weather service

The Agencia Española de Meteorología (AEMET) is the official weather service of Government of Spain, (AEMET, 2016). The service provides data about observations as well as forecasting and also historical data can be obtained. Alerts from extreme meteorological conditions (wind, rain, snow) can be retrieved from the web site, represented over a map of Spain terrirory. In the Aemet website there are forecasting data available to download in XML format. These data include forecast for 5 days and include these parameters: temperature, humidity, rain probability and sky cover index. The forecasting values are delivered for 6 hours periods.

Data are available to download in XML format via an API, example of forecasting data for Malaga city can be found here:
`http://www.aemet.es/xml/municipios/localidad_29067.xml`.

Each meteorological station has a numeric code for identification purposes, this code is used in the URL for retrieveing data. The code for the station used here is 29067.

Here is a portion of the XML response for the previous url, this XML response is related to Málaga city and to the next day weather:

```
<dia fecha="2016-01-08">
```

Figure 3.9: AEMET Website

```
<prob_precipitacion periodo="00-24">85</prob_precipitacion>
<prob_precipitacion periodo="00-12">0</prob_precipitacion>
<prob_precipitacion periodo="12-24">85</prob_precipitacion>
<prob_precipitacion periodo="00-06">0</prob_precipitacion>
<prob_precipitacion periodo="06-12">0</prob_precipitacion>
<prob_precipitacion periodo="12-18">0</prob_precipitacion>
<prob_precipitacion periodo="18-24">80</prob_precipitacion>
<cota_nieve_prov periodo="00-24"/>
<cota_nieve_prov periodo="00-12"/>
<cota_nieve_prov periodo="12-24"/>
<cota_nieve_prov periodo="00-06"/>
<cota_nieve_prov periodo="06-12"/>
<cota_nieve_prov periodo="12-18"/>
<cota_nieve_prov periodo="18-24"/>
<estado_cielo periodo="00-24" descripcion="Intervalos_nubosos_con_lluvia_escasa">43</
    ↪ estado_cielo>
<estado_cielo periodo="00-12" descripcion="Despejado">11</estado_cielo>
<estado_cielo periodo="12-24" descripcion="Intervalos_nubosos_con_lluvia_escasa">43</
    ↪ estado_cielo>
<estado_cielo periodo="00-06" descripcion="Despejado">11n</estado_cielo>
<estado_cielo periodo="06-12" descripcion="Despejado">11</estado_cielo>
<estado_cielo periodo="12-18" descripcion="Poco_nuboso">12</estado_cielo>
<estado_cielo periodo="18-24" descripcion="Nuboso_con_lluvia_escasa">44n</estado_cielo>
<viento periodo="00-24">
<direccion>SE</direccion>
<velocidad>15</velocidad>
</viento>
<viento periodo="00-12">
<direccion>O</direccion>
<velocidad>15</velocidad>
</viento>
<viento periodo="12-24">
<direccion>SE</direccion>
<velocidad>15</velocidad>
</viento>
<viento periodo="00-06">
<direccion>C</direccion>
<velocidad>0</velocidad>
</viento>
<viento periodo="06-12">
<direccion>SE</direccion>
<velocidad>15</velocidad>
</viento>
<viento periodo="12-18">
<direccion>SE</direccion>
<velocidad>10</velocidad>
</viento>
<viento periodo="18-24">
<direccion>S</direccion>
<velocidad>10</velocidad>
</viento>
<racha_max periodo="00-24"/>
<racha_max periodo="00-12"/>
<racha_max periodo="12-24"/>
```

```
<racha_max periodo=”00−06”/>
<racha_max periodo=”06−12”/>
<racha_max periodo=”12−18”/>
<racha_max periodo=”18−24”/>
<temperatura>
<maxima>20</maxima>
<minima>11</minima>
<dato hora=”06”>11</dato>
<dato hora=”12”>19</dato>
<dato hora=”18”>17</dato>
<dato hora=”24”>16</dato>
</temperatura>
<sens_termica>
<maxima>20</maxima>
<minima>11</minima>
<dato hora=”06”>11</dato>
<dato hora=”12”>19</dato>
<dato hora=”18”>17</dato>
<dato hora=”24”>16</dato>
</sens_termica>
<humedad_relativa>
<maxima>90</maxima>
<minima>60</minima>
<dato hora=”06”>80</dato>
<dato hora=”12”>60</dato>
<dato hora=”18”>75</dato>
<dato hora=”24”>85</dato>
</humedad_relativa>
<uv_max>2</uv_max>
</dia>
```

For this study only next day forecasted data are used because the desired forecasting horizon is 24 hours. For the next day these are the meteorological data parameters available from AEMET:

- Rain probability: is an estimation of the probability that it will rain in that period of time (%).
- Snow line: the altitude above which snow and ice can be found (meters).
- Sky cover: Sky cover indicates the amount of clouds in the sky.
- Wind speed and direction: direction (N,S,E,W) and speed (km/h).
- Air temperature: is the ambient air temperature expressed in °C.
- Humidity: is the usual relative air humidity (%).

These data are provided in a 6 hours period basis (0 to 6, 6 to 12, 12 to 18 and 18 to 24 h.).

A data base to store these data has been developed because past data are not avaiable from AEMET and these data are needed to train any algorithm or system to forecast some variables. The data base has been developed over PostgreSQL 9.0.

These data is read from AEMET and stored in the data base every day, although not all fields are stored, but only the interesting ones: air temperature, humidity, rain probability and sky cover. Available sky cover levels are shown in table 3.1.

Code	Description
11	Clear
12	Little Cloud
13	Cloudy intervals
14	Cloud
15	Cloudy
16	Covered
17	High clouds
23	Cloudy intervals with rain
24	Cloud with rain
25	Cloudy with rain
26	Covered with rain
35	Cloudy with snow
43	Cloudy intervals and low rainfall
44	Cloud with a little rain
45	Cloudy with a little rain
46	Covered with little rain
51	Cloudy intervals with storm
52	Cloud with storm
53	Cloudy with storm
54	Covered with storm
61	Cloud intervals with storm and low rainfall
62	Cloud with storm and low rainfall
63	Cloudy woth storm and low rainfall
64	Covered with storm and low rainfall
72	Cloud with little snow
73	Cloudy with little snow
74	Covered with little snow

Table 3.1: Sky cover codes and descriptions from AEMET

3.6 Conclusions

In this chapter statistical and data mining models that will be used in characterizacion and modeling the hourly solar global radiation and the different metrics to evaluate the performance of proposed models are presented. Moreover, the different sources of data used in this work are described. The data used are both historical record from weather public services as the AEMET (Agencia Española de Meteorología) and predictions.

In the case of forecasted data, the API developed for obtaining these data are presented. All these software will be included in the prediction of solar radation tool developed in the frame of this thesis.

Bibliography

AEMET, 2016. Agencia española de meteorología (web site). http://www.aemet.es, accessed: 2015-09-30.

Berthold, M. R., Hand, D. J. (Eds.), 2003. Intelligent Data Analysis: An Introduction, 2nd Edition. Springer Verlag.

B.Viorel, 2008. Modeling Solar Radiation at the Earths Surface. Recent Advances. Springer.

Celebi, M. E., Kingravi, H. A., Vela, P. A., 2013. A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications 40 (1), 200 – 210.
URL http://www.sciencedirect.com/science/article/pii/S0957417412008767

Chang, C. C., b. Lin, C., Sept 2001. Training v-support vector classifiers: Theory and algorithms. Neural Computation 13 (9), 2119–2147.

Chang, C. C., b. Lin, C., Aug 2002. Training v-support vector regression: Theory and algorithms. Neural Computation 14 (8), 1959–1977.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen, Z., Yang, Y., 2004. Assessing forecast accuracy measures.

Gorunescu, F., 2011. Data Mining: Concepts, Models and Techniques. Intelligent Systems Reference Library. Springer Berlin Heidelberg.

Han, J., Kamber, M., 2000. Data mining: Concepts and techniques.

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

MacQueen, J. B., 1967. Some methods for classification and analysis of multivariate observations. In: Fifth symposium on math, statistics, and probability. Berkeley, CA: University of California Press. p. 281–297.

Mirkin, B., 2012. Clustering, A Data Recovery Approach. Chapman and Hall.

Mubiru, J., Jan. 2011. Using artificial neural networks to predict direct solar irradiation. Adv. Artif. Neu. Sys. 2011, 12:12–12:12.
URL `http://dx.doi.org/10.1155/2011/142054`

OpenWeatherMap, 2016. Openweathermap web site. `http://openweathermap.org`, Accessed: 2015-07-10.

Rohatgi, V., Saleh, A., 2001. An Introduction to Probability and Statistics(2nd Ed.). Wiley-Interscience.

Rokach, L., Maimon, O. Z., 2008. Data mining with decision trees: theroy and applications.

Voyant, C., Soubdhan, T., Lauret, P., David, M., Muselli, M., 2015. Statistical parameters as a means to a priori assess the accuracy of solar forecasting models. Energy 90, Part 1, 671 – 679.
URL `http://www.sciencedirect.com/science/article/pii/S0360544215009846`

Chapter 4

Fundamentals of Solar Global Radiation

4.1 Introduction

In this chapter the different definitions and methods related to solar global radiation are analyzed. The expressions necessary to estimate the parameters related to the Sun-Earth geometry are presented. These expression will be helpful in determing how much solar irradiance is received outside the atmosphere and will be used in the characterization of hourly global radiation. Specifically, in this chapter, it has been included the algorithms for calculating variables related to the incoming solar radiation that a photovoltaic system will receive, i.e.: solar angles (height,azimuth and incidence) and the clearness index. In addition, an analysis of hourly global solar radiation series is presented in order to know the main characteristics of these series.

The terms solar radiation, irradiation, radiance, irradiance, luminance and illuminance are frequently found in related literature. Next, some of these terms meanings are explained briefly. Solar radiation or luminance refers to the emanated energy from the Sun, luminance is the contained energy within the visible part of the solar radiation spectrum (0.39-0.78 µm). The terms irradiation and illumination refer to the cumulative energy incident on a surface within a given period of time. Irradiance and illuminance refers to instantaneous incident energy. Table 4.1 shows names and units for these measures.

Measure	Definition	Unit	Symbol
irradiance	power density of radiation incident on a surface	G	W/m^2
irradiation	incident energy per unit area of a surface, found by integration of irradiance over a specified time interval, often an hour or a day	H	$MJ \cdot m^{-2}$
global irradiance	hemispherical solar irradiance on a horizontal plane	-	W/m^2
global radiation	hemispherical solar radiation received by a horizontal plane	-	-
extraterrestrial solar radiation	solar radiation received at the limit of the Earth's atmosphere	-	-

Table 4.1: Measure types and units

Solar irradiance must cross the atmosphere before reaching the Earth's surface and the atmosphere is affected by some complex phenomena. It's easy to know the exact amount of energy emited by the Sun at a certain moment and the amount that is "captured" by the Earth outside the atmosphere, but at the surface solar irradiance is affected by a relative high variability.

Solar irradiance variability at ground level is due to several factors like the presence of participating gases in the atmosphere ($H_2O, O_3, etc.$), aerosols, cloud cover, solar position and other factors like locale microclimate and the timescale used. However, most solar irradiance variability can be attributed to cloud cover and solar position. Variability due to solar position is completely deterministic but variability due to clouds is mostly a stochastic process because of the clouds nature. In the south of Spain typical dust events, the so called "calima", take place, particularly in summer (Cazorla et al., 2008). This dust is an extra obstacle for the solar radiation to reach the surface.

4.2 Solar position angles

The formulas related with the position of the Sun in the celestial sphere are described in this section. These expressions will be used to determine the angle of incidence of the Sun and the clearness index. Most of the input parameters used for modeling the hourly solar global radiation are based on different formulas that describe the solar position taking into account the Earth-Sun astronomical relationships, giving the Sun's coordinates in a specific reference system.

4.2.1 Daily angle

Daily angle (Γ) is the angle that determines the Earth's position on it's orbit arround the Sun. It only depends on the day of the year and can be calculated from this equation:

$$\Gamma = 2\pi \frac{d_n - 1}{365} \tag{4.1}$$

4.2.2 Declination angle

The declination angle is the angle between the Earth's rotation axis and the perpendicular plane that crosses the Earth's center. The Spencer formula can be used to calculate the declination angle (in radians) on any given day:

$$\begin{aligned} \delta(radians) = 0,006918 - 0,399912 \cdot \cos(\Gamma) + 0,070257 \cdot \sin(\Gamma) - 0,006758 \cdot \cos(2\Gamma) \\ +0,000907 \cdot \sin(2\Gamma) - 0,002697 \cdot \cos(3\Gamma) + 0,00148 \cdot \sin(3\Gamma) \end{aligned} \tag{4.2}$$

where:
Γ = daily angle

4.2.3 Hour Angle

The hour angle (ω) is positive during the morning, reduces to zero at solar noon and becomes increasingly negative as the afternoon progresses. It can be calculated from the following equation:

$$\omega(radians) = \frac{2\pi}{24}(TST(\Phi) - 12) \tag{4.3}$$

where:
Φ = latitude of the observer

Figure 4.1 shows the solar angle ω over the Earth's globe.

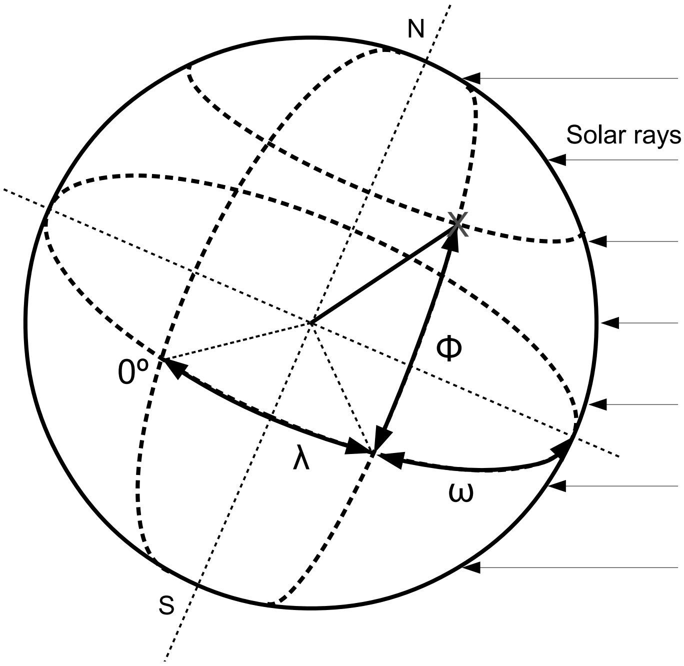


Figure 4.1: Latitude Φ, longitude λ and hour angle ω

4.2.4 Azimuth angle

Let be the imaginary line that connects the Sun with our position and let be the projection of this line on the Earth's surface. The angle of this line with the geographical south line is called the azimuth angle. At dawn it has a negative value (east), it's zero at solar noon and becomes positive at evening.

The azimuth angle (Ψ)can be calculated from the following equation:

$$\cos(\Psi) = \frac{\sin(e) \cdot \sin(\Phi) - \sin(\delta)}{\cos(e) \cdot \cos(\Phi)} \tag{4.4}$$

where:
Ψ = azimuth angle
e = elevation angle
Φ = latitude
δ = declination angle

Figure 4.2 shows the azimuth angle Ψ from an observer point of view.

4.2.5 Zenith angle and elevation

The zenith angle or angle of incidence of the Sun on an horizontal surface (θ) and the solar elevation e can be calculated from:

$$\cos(\theta) = \sin(e) = \sin(\delta) \cdot \sin(\Phi) + \cos(\delta) \cdot \cos(\Phi) \cdot \cos(\omega) \tag{4.5}$$

where:
θ = zenith angle
e = elevation angle
δ = declination angle
Φ = latitude
ω = hourly angle

The elevation (e) angle is the angle between the line that connects the Sun with our position and it's projection on the Earth's surface and it's the complementary of the zenith angle. Figure 4.2 shows the zenith angle θ and elevation e from an observer point of view.

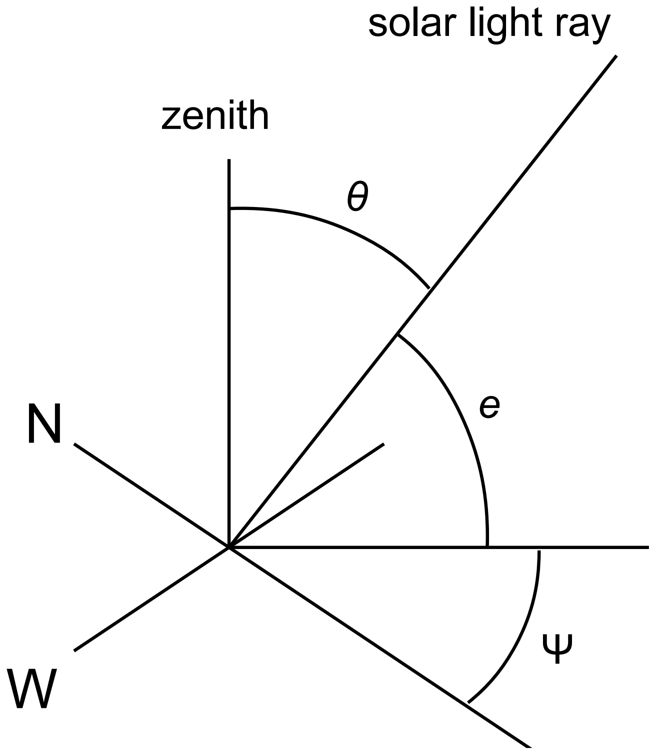


Figure 4.2: Azimuth angle Ψ, solar elevation e and zenith angle θ

4.3 Extraterrestrial solar radiation

The extraterrestrial radiation is the radiation received outside the atmosphere or the radiation at the surface if there were not atmosphere. It can be calculated using the solar constant and the exact distance between Sun and Earth. The *solar constant* is defined as the amount of total solar energy (including all wavelengths) received by unit time and unit area at the mean Sun-Earth distance. The *solar constant* was considered effectively a constant but nowadays is recognised that this term is varying continuosly, from minutes to years or even decades, but particularly within the 11-years solar activity cycle. Therefore, this magnitude is referred to as *total solar irradiance* (TSI) and the term *solar constant* can be used to describe the long-term average of TSI. For I_{sc}, (Gueymard, 2004) gives a value of 1366.1 Wm^{-2}, years before (Iqbal, 1983) gave a value of 1367 Wm_{-2}

The extraterrestrial radiation can be calculated using the folloging formulas. The hourly extraterrestrial radiation, $G_{0,h}$, received on an horizontal surface is obtained using equation 4.6:

$$G_{0,h} = I_{sc}E_0 \cos\theta_z = I_{sc}E_0(\sin\delta\sin\phi + \cos\delta\cos\phi\cos\omega_s)(Whm^{-2}), \qquad (4.6)$$

where I_{sc} is the solar constant, E_0 is the eccentricity factor, δ is the declination angle, ϕ is the latitude, ω_s is the hour angle centered at $(\omega_s - \pi/24, \omega_s + \pi/24)$. The expressions to estimate E_0, δ and ω_s can be found in (Iqbal, 1983).

The daily extraterrestrial radiation, $G_{0,d}$, received on an horizontal surface is obtained using equation 4.7:

$$G_{0,d} = \frac{24}{\pi} I_{sc}E_0(\omega_{sr}\sin\delta\sin\phi + \cos\delta\cos\phi\sin\omega_{sr})(Whm^{-2}), \qquad (4.7)$$

where ω_{sr} is sunrise angle.

4.4 Hourly series of global solar radiation

Hourly series are used as base to study and analyze solar radiation characteristics and behaviour and as input to the machine learning algorithms in order to forecast solar global radiation.

Data recorded by meteorological stations can be recorded at different periods of time, for exaple, every minute, every 10 minutes, every hour, etc. In this case, available data are recorded usually every minute so previous processing is necessary in order to obtain hourly series from it.

4.4.1 Analysis of series

It's desirable that data series (instantaneous or hourly) show no trends (seasonal or daily) so the prediction model can concentrate on the stochastic behaviour

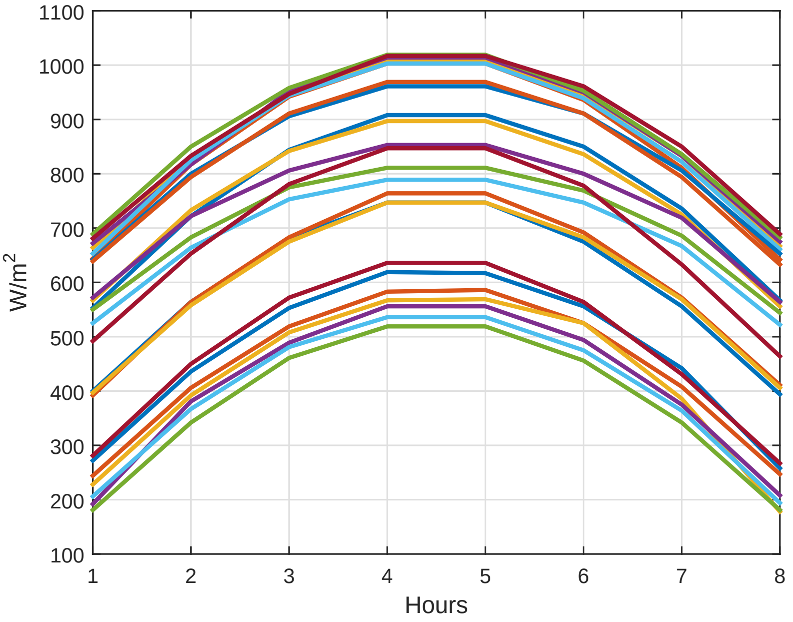


Figure 4.3: Hourly global radiation for several days of year 2013

of the series. Series that avoid showing these trends would be better as input for forecasting methods because these forecasting methods would not have to predict different radiation levels in summer and winter but only manage to predict solar radiation values according to weather conditions.

The hourly solar extraterrestrial and solar global radiation at noon throughout the year 2013 are shown in figure 4.4, where the trends in summer and winter can be easily noted.

Figure 4.3 shows the global solar horizontal radiation for several day at the same station. It's clear that along the different year seasons radiation levels can be very different.

It's desirable that for the series to be stationary. Thus, study series (hourly and daily) of global solar radiation can be done following these steps:

- Find and characterize observed trends and eliminate them from original series.

- Analyze statistical properties of the resulting series: mean values, variance,

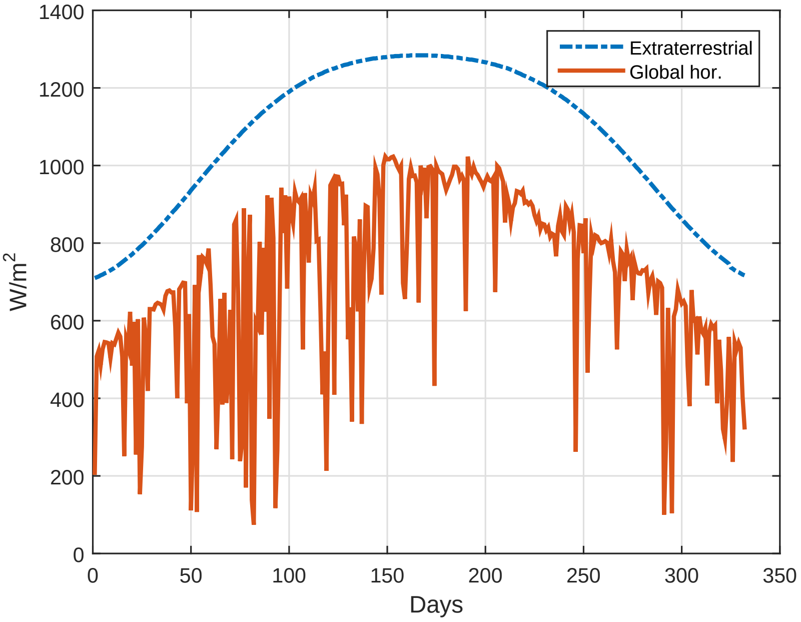


Figure 4.4: Hourly global radiation at 12h and extraterrestrial hourly global radiation (left) and hourly clearness index at 12 h (right) for year 2013

distribution functions, etc.

- Depending on observed properties, propose the best models to simulate original series and predict new values.

Next, some techniques to eliminate daily and annual trends are reviewed.

4.4.2 Methods to remove the seasonal and daily trend in global radiation series

In the literature there are several ways to eliminate seasonal and daily trends for solar global radiaton, some of them are:

- Average or moving average using Fourier functions
- Physical models
- Maximum exposure hourly global radiation values

Average or moving average using Fourier functions

In (Balouktsis and Tsalides, 1986) authors obtain Fourier coefficients for each hour of the day along the year and use this function:

$$G^F_{h,d} = A_{0,h} + A_{1,h} \cos \frac{2\pi(d-1)}{L} + B_{1,h} \sin \frac{2\pi(d-1)}{L} \tag{4.8}$$

where $G^F_{h,d}$ is the value of the Fourier series for day d and hour h; B_i, A_i are the Fourier coefficients and L is the considered hour period. Using these $G^F_{h,d}$ values, an index F is calculated as follows:

$$F_{h,d} = \frac{G_{h,d}}{G^F_{h,d}} \tag{4.9}$$

However, this series is not normal and it is necessary to use gaussin techniques.

Physical models

A "clear sky" model is a model capable of estimate the solar irradiance received at the Earth's surface. It may produce estimates of direct (E_{bnc}), diffuse (E_{dhc}) and global (E_{ghc}) solar irradiance (Engerer and Mills, 2015).

For example, (Ineichen and Perez, 2002) describes a model for direct and global radiation as:

$$\begin{aligned} E_{bnc} =& b * E_{ext_n} * \exp(-0.09 * AM * (T_L - 1)) \\ E_{ghc} =& a_1 * E_{ext_n} * \cos(\Theta_z) * \exp(-a_2 * AM) * (f_{h1} + f_{h2} * (T_L - 1)) \end{aligned} \tag{4.10}$$

These formulations include empirical adjustements for altitude coefficients and incorporate also turbidity information via Linke turbidity coefficient.

(Ineichen, 2008) proposed a simplified version of the Solis model, which was computationally expensive and required sparsely measured input. The simplified model is easier to use and is capable of producing estimates for direct, diffuse and global radiation:

$$\begin{aligned}
E_{bnc} =& E'_{ext} * \exp(-\tau_b / \cos(\Theta_z)^b) \\
E_{ghc} =& E'_{ext} * \exp(-\tau_g / \cos(\Theta_z)^g) * \cos(\Theta_z) \\
E_{dhc} =& E'_{ext} * \exp(-\tau_d / \cos(\Theta_z)^d)
\end{aligned} \tag{4.11}$$

where $E'_{ext}, \tau_b, \tau_g, \tau_d$ are all dependent on the aerosol optical depth, water vapor and atmospheric pressure are also required as inputs.

The Esra Model was developed for the Europan Solar Radiation Atlas (Rigollier et al., 2000). The model requires the Linke turbidity as input and uses and air mass based parametrization for Rayleigh optical thickness (δ_R). The model can estimate direct and diffuse solar radiation:

$$\begin{aligned}
E_{bnc} =& E_{ext_n} * \exp(-0.8662 * T_L * AM * \delta_R) \\
E_{dhc} =& E_{ext_n} * T_{Rd}(T_L) * F_d(\Theta_z, T_L)
\end{aligned} \tag{4.12}$$

The REST2 model (Gueymard, 2008) is separated into two bands representing the broadband components of two separate series of spectra and incorporates transmission estimates for Rayleigh scattering (T_{R_i}), uniform gas (T_{G_i}), ozone (T_{O_i}), nitrogen dioxide (T_{N_i}), and water vapor (T_{W_i}) absorption and aerosol extinction (T_{A_i}). The beam estimate is:

$$E_{bnc_i} = E_{ext_{ni}} * T_{R_i} * T_{G_i} * T_{O_i} * T_{N_i} * T_{W_i} * T_{A_i} \tag{4.13}$$

The diffuse clear sky estimate (E_{d_i}) is divided into two layers, in the upper layer component (E_{dp_i}) Rayleigh scattering, ozone and uniform gas absorption are estimated. For the bottom layer (E_{dd_i}), aerosol, water vapor, and nitrogen dioxide absorption and, separately, aerosol scattering (T_{asi}), ground and sky albedo (ρ_{G_i}, ρ_{S_i}) are used:

$$\begin{aligned}
E_{dp_i} =& E_{ext_{hi}} * T_{G_i} * T_{O_i} * T_{N_i} * T_{W_i} \\
& * [B_{R_i} * (1 - *T_{R_i}) * T_{A_i}^{0.25} + B_A * F_i * T_{R_i} * (1 - *T_{asi}^{0.25})] \\
E_{dd_i} =& \rho_{G_i} * \rho_{S_i} * (E_{bn_i} * \cos(\Theta_z) + E_{dp_i})/(1 - \rho_{G_i} * \rho_{S_i})
\end{aligned} \tag{4.14}$$

Maximum exposure hourly global radiation values

This method is based on the maximum expected values of hourly exposition and can be also used for daily exposition series. (Boileau, 1983) describes two models for forecasting solar radiation, with first, Model B, it studies the increments of daily global radiation from one day to the next and shows that the fluctuations can be modeled using a white noise a_i:

$$v_i = a_i - \Theta a_{i-1} \tag{4.15}$$

It was demonstrated that the same value for parameter Θ ($\Theta = 0.75$) can be used for the three sites considered. With the second model, Model C, better results are achieved:

$$\tilde{I}_i = I_i - \langle I_i \rangle \tag{4.16}$$

so forecasts are calculated with one past day only using a Markov model with the centered variable $\tilde{I}_i$.

4.4.3 Clearness Index

Clearness index is commonly used to analyze and characterize solar global radiation as this index allows the seasonal and daily trends observed in solar radiation to be removed. Using this index instead of clear sky index is proposed as it is possible to estimate the first one only using well known expressions without knowing the specific weather conditions of the locations, sometimes necessary for some of clear sky models, (Zhong and Kleissl, 2015). The main problem of using *Clearness index* is that obtained series are not truly stationary timeseries.

The clearness index is defined as the ratio of the horizontal global radiation to the radiation received outside the atmosphere or extraterrestrial irradiance according to Expression 4.17:

$$K_t = \frac{G_t}{G_{t,0}}, \tag{4.17}$$

where G_t is the solar global radiation recorded for time t and $G_{t,0}$ is the extraterrestrial solar global radiation for this period.

According to (Woyte et al., 2007), the instantaneous *Clearness index* K_t or transmission coefficient over a arbitrarly oriented surface is:

$$K_t = \frac{G}{S_0 E \cos I} \tag{4.18}$$

where S_0 is the solar constant, E is the eccentricity correction factor and I is the angle of incidence. This formulation of K_t is dependent on the zenith angle.

4.5 Solar forecast metrics

To quantify the accuracy of solar forecasts various metrics can be used. Determining which one is most appropiate depends on the user: while system operators need metrics that accurately reflects the cost of forecast errors, researchers require indicators of relative performance of different forecast models and indicators for a single model under different conditions. Also, an appropiate test dataset and analysis procedure are very important. Test dataset must exclude all data used to train models so that evaluation is performed on new or unseen data. Data has to be checked to remove observation errors and hardware and operation issues so that forecast is done over good quality data without forecasting deterministic data.

Performance metrics can be categorized according to three types of forecasting error: bias, variance and correlation. In (Hoff et al., 2012) several absolute and relative statistical metrics for errors in forecasting are analyzed and it shows that a large number of metrics are needed to provide a clear picture of the forecasting accuracy of any method. In 3 several general error methods are introduced: MSE, $RMSE$, MAE, etc.

Bias characterizes the balance between over and underprediction. *Mean Bias Error* is the most commonly used bias measure and is defined as

$$MBE = \frac{1}{N} \sum_{t=1}^{N} (I(t) - \hat{I}(t)) \tag{4.19}$$

where $I(t)$ is the measured irradiance at time t, $\hat{I}(t)$ is the forecasted irradiance at time t and N is the number of data points in the data set. If the case of a perfect forecast ($\hat{I}(t) = I(t)$) this metric returns 0 but also for situations where positive and negative errors simply cancel out by summing to 0.

The coefficient of determination R^2 is able to measure how well forecast values predict trends in measured values, it is a comparison between the variance of the errors and the variance of the data to be modeled:

$$R^2 = 1 - \frac{\sigma^2(\hat{I} - I)}{\sigma^2(I)} \tag{4.20}$$

where σ^2 is the variance of the dataset. For perfect forecasting $R^2 = 1$. The value of R^2 is directly related to the $RMSE$ by noticing that:

$$R^2 \approx 1 - \frac{RMSE^2}{var(I)} \tag{4.21}$$

To evaluate variance of forecast errors two metrics are commonly used: the root mean square error ($RMSE$) and the mean absolute error (MAE). $RMSE$ is related to the standard deviation of the errors. $RMSE$ is calculated as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\hat{I}(t) - I(t))^2} \tag{4.22}$$

Tipically night values are removed in the above calculations of R^2 and $RMSE$.

It is usual to normalize values relative to energy produced or energy capacity; utilities tend to use the second one (more favorable) while scientists tend to prefer the first one. None of these metrics take into account the variability in the irradiance time-series data. (Perez et al., 2010) found that $RMSE$ error is lower in places with sunnier (less variable) weather conditions.

4.5.1 Persistence model

In general, a persistence model is a model that assumes that the next value will be equal to the current observed value. For example, a persistence model for the exchange rate between Euro and Dollar would predict that tomorrow the value of the exchage rate will be the same as today. Another example, a persistence model for the wind speed would say that the wind speed for the next hour will be the same speed as now.

For hourly solar irradiance, a persistence model can be built taking into account that solar irradiance varies within the day hours. It makes no sense assuming that irradiance will be the same along the different day hours. A basic model can be built assuming that for a given day, the irradiance at a certain hour will be the same to the irradiace at the previous day at the same hour.

For daily solar irradiance, a simpler model can be built assuming that the daily irradiance for the next day will be the same as for current day.

4.5.2 Forecast Skill

The forecast skill over 24 hour persistence forecasts, s, proposed in (Coimbra and Kleissl, 2013), is estimated according to equation 4.23:

$$s = 1 - \frac{RMSE_{model}}{RMSE_{persistence}}. \tag{4.23}$$

The forecast skill is independent of the specific meteorological or climatological characteristics of the site under consideration, (Coimbra and Kleissl, 2013). The forecast skill as defined above is such that when $s = 1$ the solar forecast is perfect and when $s = 0$ the forecast uncertainty is as large as the variability. By definition, the persistence model should have a forecast skill $s = 0$, because the error is equal to the variability, so, the ratio expressed in 4.23 is a measure of improvement over the persistence forecast. If the value of s is negative, then, the model performance is worse than the persistence forecast. So, forecast models with better performance than persistence model will have a value for s between 0 and 1, with higher values for better performance.

4.6 Conclusions

In this chapter, first, solar angles are presented as they can be used to calculate extraterrestrial solar radiation, which at the same time, is useful to calculate clearness index values. Clearness index is useful because it is free from seasonal and daily trends.

Next, hourly series are presented as the base to study and analyze solar radiation characteristics and behaviour, they are very useful as input to the machine learning algorithms in order to forecast solar global radiation. Some methods to remove seasonal and daily trends are presented, like physical models and clearness index.

Some solar radiation forecasting error metrics are also presented and analyzed, including MBE, R^2 and $RMSE$. They are useful for comparing performance of different forecasting models in different situations.

Finally, persistence model is presented as a naive model to compare against tested models. It is usefull to find out if models are really good and promising. The forecast skill that uses persistence error as benchmark is also proposed as a measurement of model errors as it is used to measure the performance of a forecasting model against persistence model.

Bibliography

Balouktsis, A., Tsalides, P., 1986. Stochastic simulation model of hourly total solar radiation. Solar Energy 37 (2), 119 – 126. URL http://www.sciencedirect.com/science/article/pii/0038092X86900691

Boileau, E., 1983. Use of some simple statistical models in solar meteorology. Solar Energy 30 (4), 333 – 339. URL http://www.sciencedirect.com/science/article/pii/0038092X83901871

Cazorla, A., Olmo, F. J., Alados-Arboledas, L., Jan 2008. Development of a sky imager for cloud cover assessment. J. Opt. Soc. Am. A 25 (1), 29–39. URL http://josaa.osa.org/abstract.cfm?URI=josaa-25-1-29

Coimbra, C., Kleissl, J., 2013. Solar Resource Assessment and Forecasting. Elsevier, Waltham, Massachusetts, Ch. Chapter 8: Overview of Solar-Forecasting Methods and Metric for Accuracy Evaluation, pp. 171–194.

Engerer, N., Mills, F., 2015. Validating nine clear sky radiation models in australia. Solar Energy 120, 9 – 24. URL http://www.sciencedirect.com/science/article/pii/S0038092X15003527

Gueymard, C. A., 2004. The sun's total and spectral irradiance for solar energy applications and solar radiation models. Solar Energy 76 (4), 423 – 453. URL http://www.sciencedirect.com/science/article/pii/S0038092X03003967

Gueymard, C. A., 2008. Rest2: High-performance solar radiation model for cloudless-sky irradiance, illuminance, and photosynthetically active radiation – validation with a benchmark dataset. Solar Energy 82 (3), 272 – 285. URL http://www.sciencedirect.com/science/article/pii/S0038092X07000990

Hoff, T. E., Perez, R., Kleissl, J., Renne, D., Stein, J. S., 2012. Reporting of irradiance model relative errors. In: Proceedings of the 2012 ASES Annual Conference. Rayleigh, NC.

Ineichen, P., 2008. A broadband simplified version of the solis clear sky model. Solar Energy 82 (8), 758 – 762.
URL http://www.sciencedirect.com/science/article/pii/S0038092X08000406

Ineichen, P., Perez, R., 2002. A new airmass independent formulation for the linke turbidity coefficient. Solar Energy 73 (3), 151 – 157.
URL http://www.sciencedirect.com/science/article/pii/S0038092X02000452

Iqbal, M., 1983. An introduction to solar radiation. Academic Press Inc. New York - London.

Perez, R., Kivalov, S., Schlemmer, J., Hemker, K., Renné, D., Hoff, T. E., 2010. Validation of short and medium term operational solar radiation forecasts in the us. Solar Energy 84 (12), 2161–2172.

Rigollier, C., Bauer, O., Wald, L., 2000. On the clear sky model of the {ESRA} — european solar radiation atlas — with respect to the heliosat method. Solar Energy 68 (1), 33 – 48.
URL http://www.sciencedirect.com/science/article/pii/S0038092X99000559

Woyte, A., Belmans, R., Nijs, J., 2007. Fluctuations in instantaneous clearness index: Analysis and statistics. Solar Energy 81 (2), 195 – 206.
URL http://www.sciencedirect.com/science/article/pii/S0038092X0600079X

Zhong, X., Kleissl, J., 2015. Clear sky irradiances using rest2 and modis. Solar Energy 116, 144 – 164.

Chapter 5

Forecasting hourly profiles of solar global radiation using Cumulative Probability Distribution Functions

5.1 Introduction

In this chapter a technique to model hourly profiles of solar global radiation is proposed. To address the need of forecasting hourly solar radiation values a method that proposes using statistical and data mining techniques that allow finding and estabilishing different hourly profiles of solar radiation for different days is introduced, then a new method for forecasting hourly profiles of solar radiation is proposed using clearness index.

First, clearness index is calculated for every recorded value, then CPDF are applied in order to cluster them into different groups using K-means, so there will be different groups according to different types of day. Next a cluster selection is proposed using K_d value (this may be a forecasted value) and finally, an hourly global radiation profile for a given day is calculated using the selected cluster mean (or centroid) and extraterrestrial radiation values which are easily calculated. A method to measure errors in forecasted values is also proposed.

5.2 Clearness Index

As mentioned in chapter 3 data used in this experiment ranges from 2010/11/1 to 2012/10/31 in order to have two complete years. Registers are recorded every minute and each contains the observed solar global horizontal radiation as well as the extraterrestrial solar horizontal radiation (calculated). Clearness index values (K_{inst}) are estimated for each register using equation 4.17. Once K_{inst} values are calculated, registers are summarized into 20 minutes period records, therefore obtaining K_{20min} values. Figures 5.1, 5.2 and 5.3 show the K_{inst} and K_{20min} for three different days.

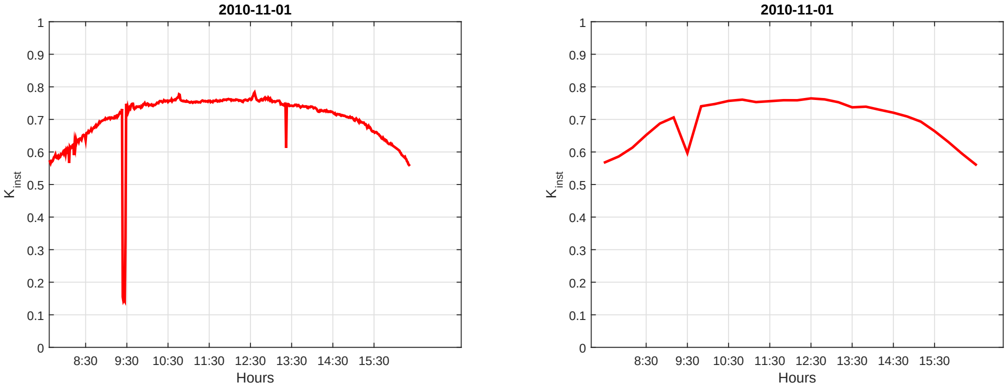


Figure 5.1: K_{inst} (left) and K_{20min} (right) for day 2010-11-01

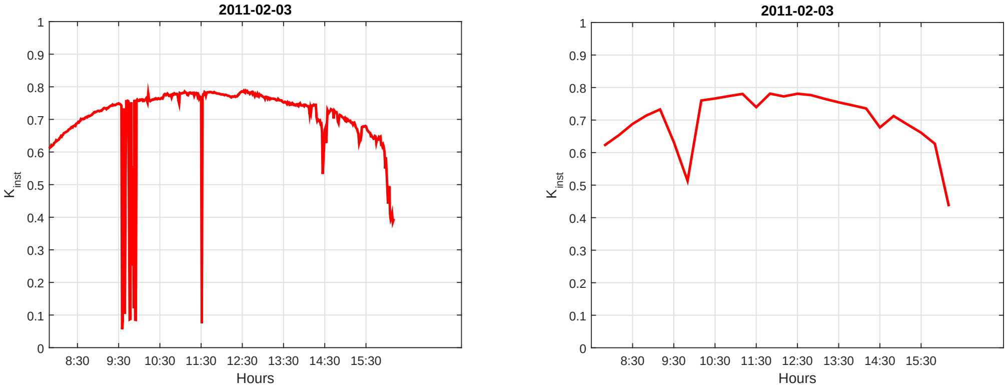


Figure 5.2: K_{inst} (left) and K_{20min} (right) for day 2011-02-03

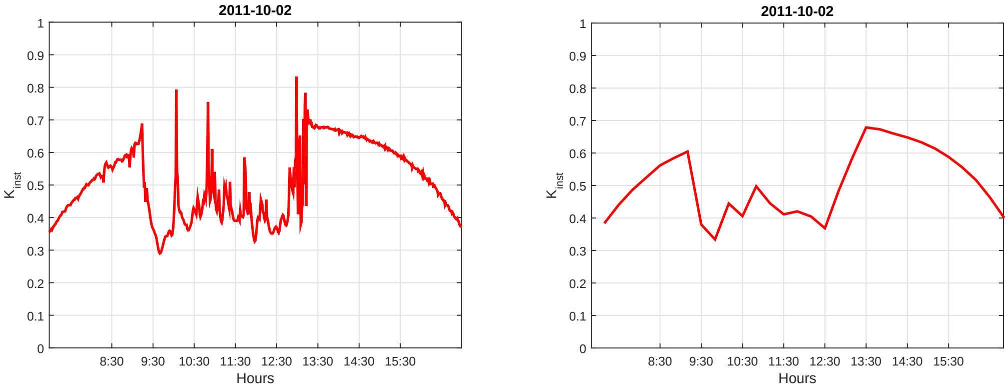


Figure 5.3: K_{inst} (left) and K_{20min} (right) for day 2011-10-02

5.3 Cumulative Probability Distribution Functions

After calculating K_{20min} values, the probability distribution functions (CPDF) are calculated for the values of each day and a vector of dimension 100 is obtained as the accuracy used for estimating clearness index is 0.01. In the first step, a vector representing the number of values found in a certain range is obtained. For example, if we are calculating CPDF wih a resolution of 100 different intervals, with values between 0 and 1, and there are 3 values between 0.51 and 0.52, that is $0.51 < X \leq 0.52$, then the value for 0.51 will be three. Left images from figures 5.4, 5.5 and 5.6 show the graphs corresponding to the vectors of this first step calculated for days 2010-11-01, 2011-02-03 and 2011-10-02 respectively.

Second step is to calculated CPDF using the previous calculated vectors. For each vector index i, the value of CPDF vector is the sum of all vector values from 1 to i. Finally, the values of the CPDF vector are normalized by dividing all values by the number of values added.

Right images from figures 5.4, 5.5 and 5.6 show the graphs corresponding to the CPDF functions calculated for days 2010-11-01, 2011-02-03 and 2011-10-02 respectively.

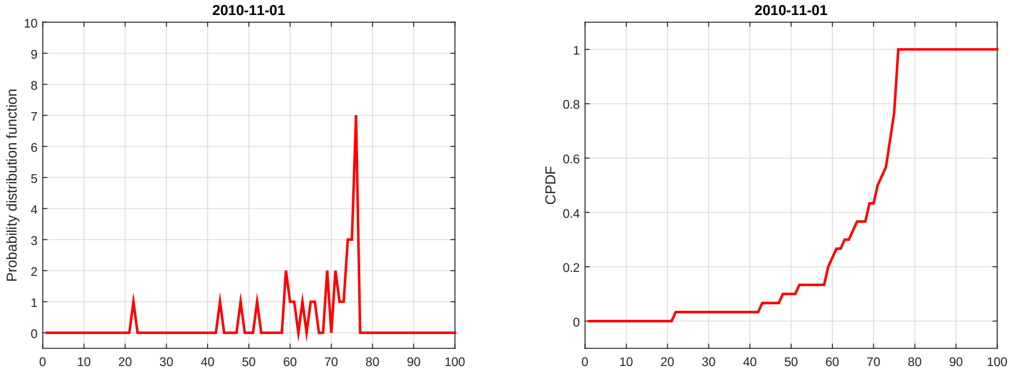


Figure 5.4: probability distribution values (left) and CPDF (right) for day 2010-11-01

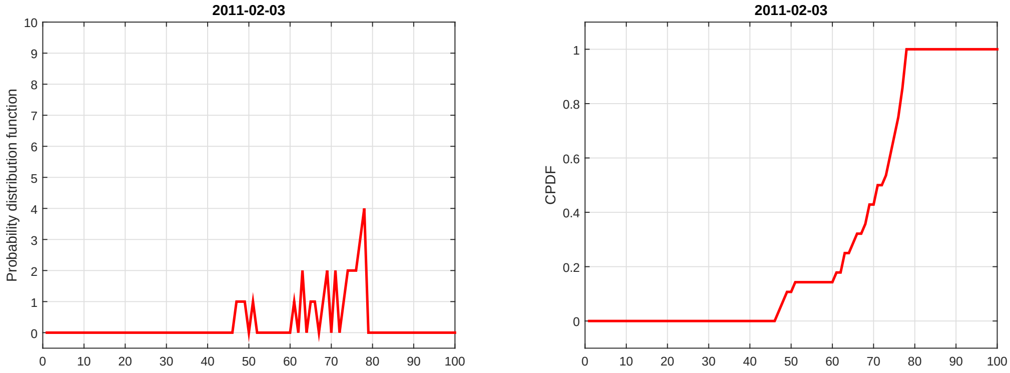


Figure 5.5: Accumulated values (left) and CPDF (right) for day 2011-02-03

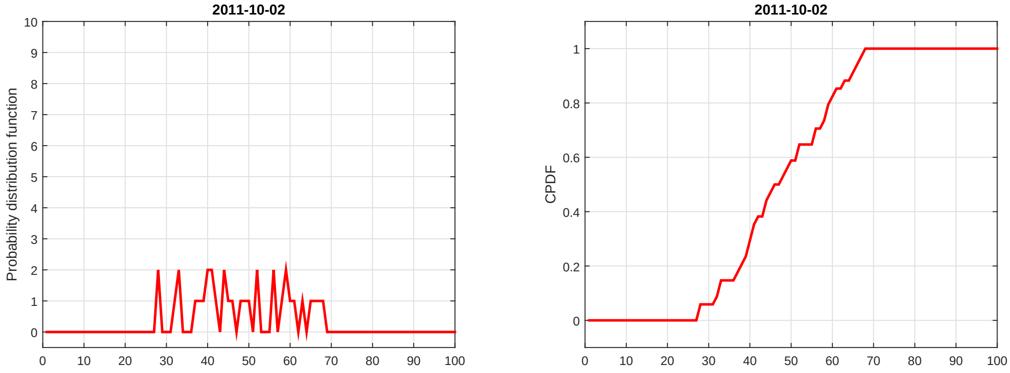


Figure 5.6: Accumulated values (left) and CPDF (right) for day 2011-10-02

5.4 Clustering CPDF

There are different Cumulative Probability Distribution Functions depending on the observed values of clearness index estimated each 20 minutes. The proposal is to obtain the minimum number of CPDF that represent all observed curves using a clustering technique that cluster them and to prove that these curves are statistically equals to the rest of estimated CPDF. To achieve this goal the use of a clustering technique is proposed. Specifically, the K-means method has been used.

CPDF vectors are used as the inputs for K-means clustering. The number of clusters need to be fixed in advance in order to use K-means. We checked four different number of clusters, from 4 to 7, taking into account the expected different CPDF's. For each execution of K-means we obtained the centroid of each cluster. Using these centroids we checked the equality of CPDF's of the cluster in each test using the Kolmogorov-Smirnov two sample test. Table 5.1 shows the results obtained for the different numbers of checked clusters when using $\alpha = 0.05$ (significance level).

Number of	CPDF's for which	% CPDF's	% of clusters
clusters	$D_{n,m} > c_\alpha$		$D_{n,m} > c_\alpha$
4	29	4.4	100
5	26	3.9	80
6	21	3.1	33
7	19	2.9	29

Table 5.1: Results of Kolmogorov-Smirnov two sample test for different number of clusters

Taking these results into account, 6 clusters are selected instead of 7 clusters as only 2 of 6 clusters in both cases have some CPDF's (approximately 3%) that are significantly different from theirs centroids and therefore 6 clusters are enough to capture the different CPDF's observed. The days included in each cluster are from different months in all cases which means that the clustering method allows us to capture the different CPDF's observed along the year and not the season of the year.

Figures 5.7, 5.8 and 5.9 shows CPDF's distribution for each cluster when using $K = 6$.

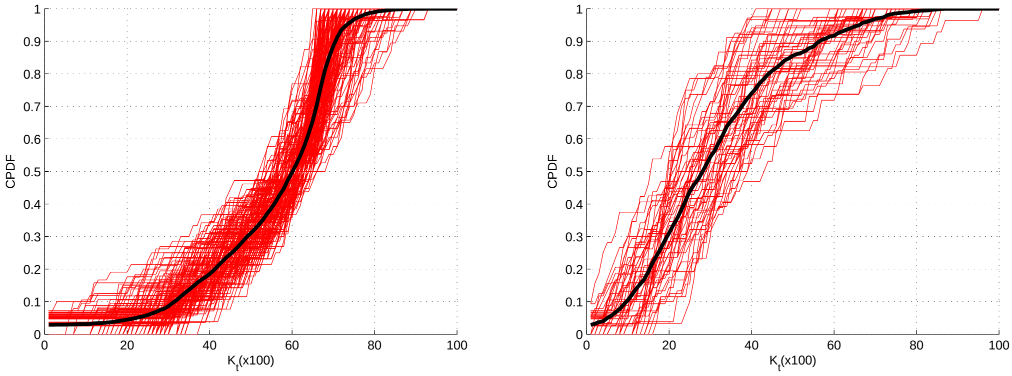


Figure 5.7: CPDF's for clusters 1 (left) and 2 (right)

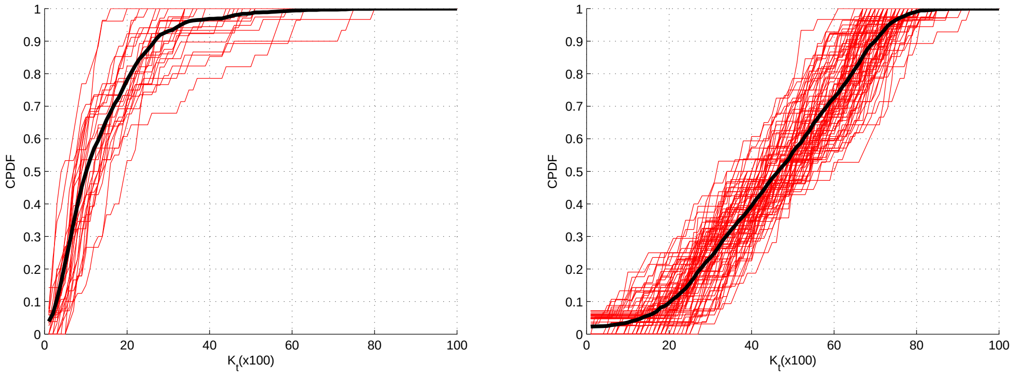


Figure 5.8: CPDF's for clusters 3 (left) and 4 (right)

5.5 Relation between K_d values and clusters

After clustering all the observations, the relationship between the daily clearness index corresponding to each CPDF and its cluster is analyzed. Figure 5.10 shows these values.

As can be observed, the daily clearness index is related to the cluster to which the CPDF for the given day belongs. These results can be used to decide the cluster to which one day belongs. It has been checked that for CPDF's that can belong to two different clusters (see Fig.5.10) the $D_{n,m}$ between these each CPDF and the centroid of each possible cluster is always lower than the critical value c_α. Assigning the cluster depending on the daily clearness index value using the expression 5.1 is proposed.

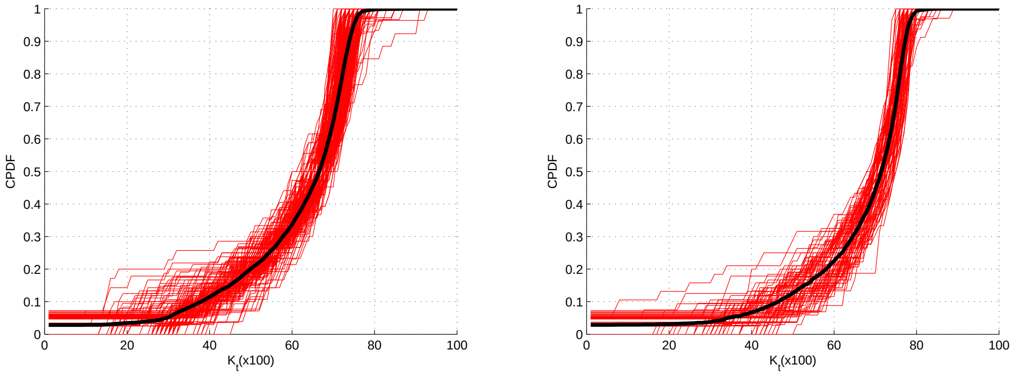


Figure 5.9: CPDF's for clusters 5 (left) and 6 (right)

$$
Number\ of\ cluster = \begin{cases} 3 & \text{if } K_d \leq 0.22 \\ 2 & \text{if } 0.22 < K_d \leq 0.42 \\ 4 & \text{if } 0.42 < K_d \leq 0.55 \\ 1 & \text{if } 0.55 < K_d \leq 0.62 \\ 5 & \text{if } 0.62 < K_d \leq 0.7 \\ 6 & \text{if } K_d > 0.7 \end{cases} \tag{5.1}
$$

The observed relationship between daily clearness index and cluster suggests that K-means could produce clusters with days with a similar hourly solar radiation profile. Moreover, the relationship between the solar radiation distribution during a day and the daily clearness index value has been pointed out in (Bendt et al., 1981). However, solely one hourly profile cannot be used for all days due to the observed differences in different clusters. Using these two facts, using all the days of each cluster to estimate the solar radiation hourly profile for that cluster is proposed. Therefore, the hourly clearness index mean value and its standard deviation have been calculated for each cluster. Figures 5.11, 5.12 and 5.13 show these values. As can be observed, the standard deviation for most clusters and hours is not large and the results show that these values change with solar time, particularly for the hours at the start and end of the day. These results agree with those previously obtained in (Aguiar and Collares-Pereira, 1992). Moreover, these values decrease significantly for the hours with more radiation (central hours of day). This is the reason for proposing the use of these mean values as the hourly profile model for each cluster.

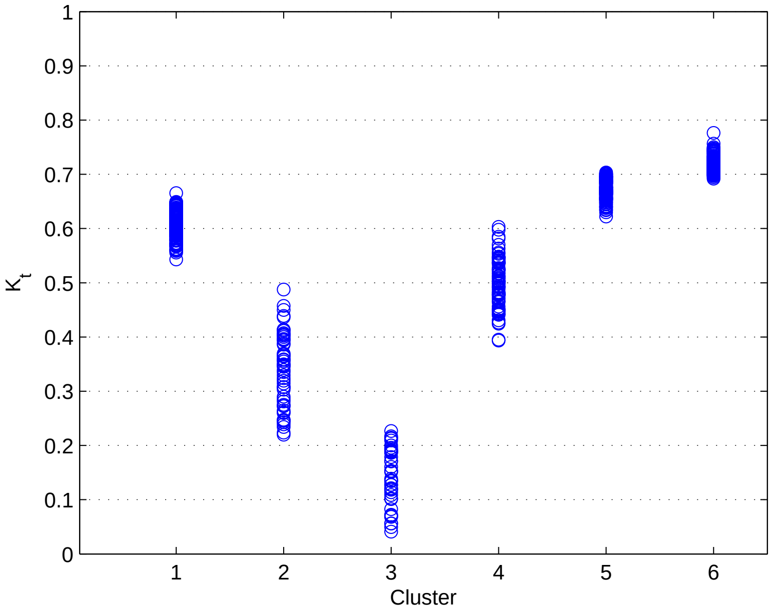


Figure 5.10: Daily clearness index, K_d, vs cluster to which the day belongs

5.6 Forecasting

With the daily clearness index and the selected cluster obtained using equation 5.1, the hourly K_h profile of the cluster can be selected and used to forecast the hourly radiation values simply by multiplying each hourly K_h value by its corresponding extraterrestrial radiation value, that are well known values because they are deterministic, only depending on the Sun position relative to Earth. This process is described in algorithm 3.

```
Input  : $K_d$ Daily clearness index; clearness index hourly profiles.
Using $K_d$ select the cluster.
Estimate solar global radiation hourly values using the clearness index
hourly profile for the selected cluster and the extraterrestrial solar global
radiation hourly values (Eq.4.6).
Output: Hourly solar radiation values for day $d$
```

Algorithm 3: Procedure for obtaining solar global radiation hourly values.

Following the procedure described in algorithm 3, the estimated hourly radi-

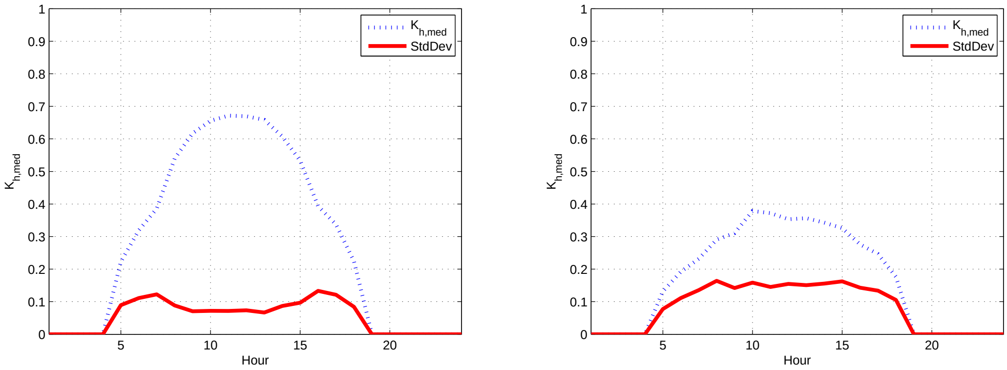


Figure 5.11: Mean hourly clearness index and hourly standard deviation for clusters 1 (left) and 2 (right)

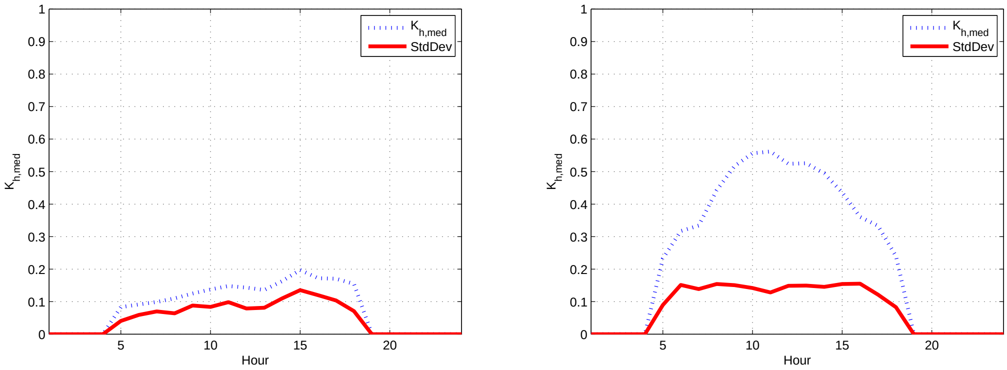


Figure 5.12: Mean hourly clearness index and hourly standard deviation for clusters 3 (left) and 4 (right)

ation values for all the recorded data are estimated for the testing dataset. The metrics for evaluating the hourly estimated modeled profiles are:

- The standard deviation of daily profile hourly values for each cluster.
- The difference between the total hourly solar radiation estimated, $\hat{G}_h$, and the total hourly solar radiation received for the whole period of data normalized to the total hourly solar radiation received, G_h, according to the expression:

$$Error_{rad} = \frac{\sum_{i=1}^{m} |G_{h,i} - \hat{G}_{h,i}|}{\sum_{i=1}^{m} G_{h,i}} 100(\%) \tag{5.2}$$

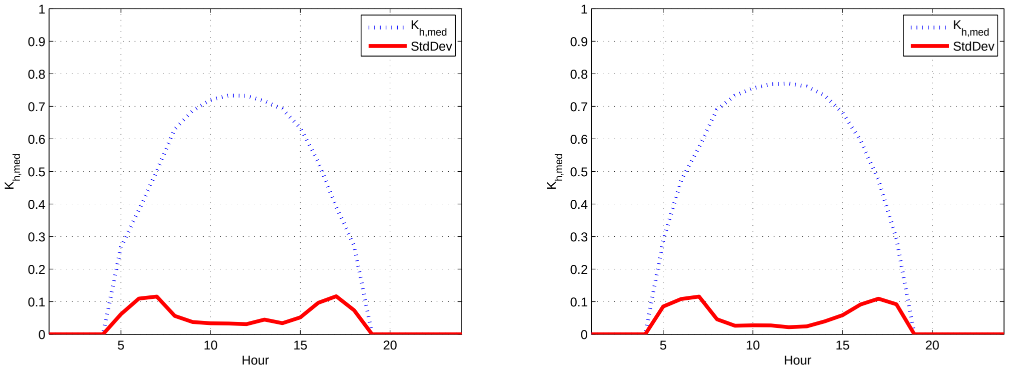


Figure 5.13: Mean hourly clearness index and hourly standard deviation for clusters 5 (left) and 6 (right)

Eq.5.2 is used to estimate the energy error for each cluster to check the accuracy of these predictions. A naive persistent model that assumes that the hourly profile for a day is the same as the profile for the previous day is also used in order to evaluate whether the proposed model improves this naive model. Table 5.2 shows the obtained results.

	EE proposed	EE naive		EE random	
Cluster	model (%)	model (%)	% energy	clustering* (%)	% energy
1	10.5	16.5	26.3	24.9	15.5
2	36.8	71.0	4.9	29.7	15.7
3	49.1	193.8	1.3	25.3	15.4
4	25.0	35.9	10.5	24.2	17.0
5	5.0	10.8	37.1	17.9	17.1
6	4.0	12.3	19.9	22.6	19.3
All clusters	10.5	20.6		24.0	

Table 5.2: Energy error (EE)(%) for each cluster when forecasting hourly solar global radiation with proposed model and with a persistent naive model and percentage of energy received in all days included in each cluster respect to total energy received.(*clusters are randomly built)

As can be observed, the proposed method is able to estimate the daily profiles of hourly global radiation with an error lower or equal to 5% for the 57% of energy received; this total increases to 84% with an error less than 11%. The highest error occurs for only 1.3% of energy received. The total error for all the clusters is 10.5% that is less than both naive model errors, 20.6%, and errors reported in previous works that range between 20 and 40%, as set out in the Introduction section. These results indicate that it is possible to use the obtained daily profiles of hourly solar radiation distribution to forecast the hourly values of this variable.

5.7 Conclusions

In this chapter a model to forecast hourly global radiation is presented. This model is based on clearness index, CPDF functions and K-means algorithm. Clearness index values are calculated and then CPDF funcions are constructed for every day in the data set. Then CPDF functions are clustered using K-means to create groups of similar days (days with similar solar radiation profile). As explaied in section 5.5, a relation between daily K_d and clusters can be established and this relation can be used to forecast hourly global radiation with a low error (5% for the 57% of energy received and 11% for the 84%).

Bibliography

Aguiar, R., Collares-Pereira, M., 1992. T.a.g: A time dependent autoregressive gaussian model for generating synthetic hourly radiation. Solar Energy 49(3), 167–174.

Bendt, P., Collares-Pereira, M., Rabl, A., 1981. The frequency distribution of daily insolation values. Solar Energy 27, 1–5.

Chapter 6

Modeling and short-term forecasting of solar global radiation

6.1 Introduction

The aim of this chapter is to explore the use of different technologies to improve accuracy in the prediction of short-term hourly global solar radiation taking into account all previous results. A new procedure to forecast next-day hourly solar global radiation is proposed. In the previous chapter, a model to estimate hourly profiles of radiation for a day was proposed. This previous model used the daily clearness index as input and was built using the cumulative probability distribution function of hourly clearness index for each day and several types of days were identified.

The proposed procedure presented here extends the previous results and is able to estimate not only the values of hourly global solar radiation for a day but also the value of the clearness index for the next day. In addition, fewer types of different days are now proposed, thanks to the use of a new variable instead of using cumulative probability distribution functions. This procedure is based on using several data mining techniques and has been checked for recorded data in Málaga, Spain.

A procedure to simulate the hourly global solar radiation for a day using different meteorological parameters (usually public data recorded by the weather services) is proposed. These simulated data can be used for the evaluation of the performance of photovoltaic facilities that do not have monitoring systems.

The system works in two different phases and applies different data mining techniques in each phase. A clustering algorithm to identify the types of days is proposed. The use of decision trees, artificial neural networks and support vector machines to estimate the parameters that characterize each type of day is also advanced. Two procedures are put forward to analize data and estimate models. Two different input data sets were used and the corresponding errors for each case are presented.

6.2 Proposed models

The proposed models are based on the following hypothesis:

- There is a limited number of day types that differ as to how hourly radiation values are distributed.
- The type of day depends on the daily clearness index value.
- The daily clearness index value is related to meteorological variables from the previous day and the current day.

Starting from these assumptions, a methodology for analyzing the recorded data is proposed in order to obtain models to forecast and simulate hourly global solar radiation values for the next day.

The data analysis is performed according to the following scheme:

1. Each observation consists of the hourly values of global radiation for a day. As these values exhibit a daily trend, a new hourly variable is proposed to remove this trend. This variable is obtained from hourly and daily clearness index.
2. Clustering the observations that have a similar daily distribution of the proposed variable and obtaining the centroid for each cluster.

3. Evaluating if it is possible to assume that the distribution of hourly values for each day is similar to the centroid of its cluster.

4. Analyzing which independent variables allows us to predict the cluster for each day and the daily clearness index using several data mining models.

6.2.1 Daily-detrended hourly clearness index

Clearness index depends on values of global horizontal irradiance. The values for a day of this index are related with the total daily radiation received and therefore with the daily clearness index. For instance, the hourly clearness indices for two days are shown in Figure 6.1 (left).

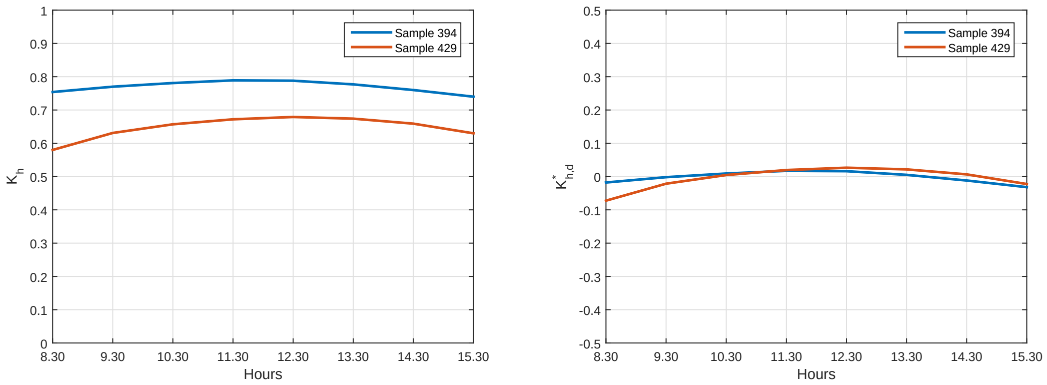


Figure 6.1: Daily profiles of hourly values of clearness index (left) and hourly values of $k^*_{h,d}$ (right) for two different days

As it can be observed, these days have a similar shape although the values of one of the curves are greater; these days correspond to days in which the changes in the clearness index are similar. To remove this dependence, a daily-detrended hourly clearness index is proposed according to:

$$k^*_{h,d} = (k_{h,d} - k_d) \qquad \texttt{for } h = 8 \texttt{ to } 15, \tag{6.1}$$

where d means day, h hour and k is the clearness index. Figure 6.1 (right) shows the values of $k^*_{h,d}$ for the same days as Figure 6.1(left). As can be observed, the new curves have both similar shape and similar values.

6.2.2 Hourly solar radiation modeling

The first step for modeling hourly solar radiation is to cluster days. The analysis process starts by estimating how many different types of day there are, while considering hourly radiation profiles throughout the day. The objective is to cluster days that have a similar profile in the same group or cluster. The K-means algorithm is used for this analysis as the hypothesis is that each one of the clusters obtained with K-means represents a different type of day.

The eight $k^*_{h,d}$ values of a day estimated using Eq. 6.1 are one observation (sample) for K-means. K-means allows us to perform the clustering in an automatic way once the number of groups (type of days) is decided. The algorithm clusters the most similar observations in each group. We propose to use $K = 4$ clusters assuming that there are 4 types of days: clear (cloudless), overcast (variable among the day), overcast mainly in the morning and overcast mainly in the afternoon. For K-means, the euclidean distance is used. As an output of K-means, four centroids are obtained and all the observations (each observation corresponds to the hourly values of one day) are clustered in one of the four clusters.

Using the type of day (that is, the corresponding centroid) and the value of daily clearness index (to obtain the hourly clearness index from daily-detrended index values from the centroid) is proposed in order to obtain the values of hourly global radiation for a day. That is, modeling and predicting hourly values of solar radiation using centroids estimated with K-means require (i) estimating the daily clearness index and (ii) the type of day (cluster). The goal, once centroids are built, is to propose models that allows us to estimate the cluster (and therefore centroid) and daily clearness index for a day. These models use the meteorological variables described in section 3.5.1 as input (independent) variables and different data mining methods are used for those two tasks.

Two different procedures are proposed for modeling the two dependent variables: in the first one, the estimation of dependent variables (cluster and daily clearness index) are modeled separately; in the second one, the variable *cluster* for each day is predicted using all data but the other independent variable, daily clearness index, is predicted in each cluster only using the observations of that cluster. Therefore, in the second procedure, one different model has been fitted for each cluster. The proposed processes for the selection of models are shown in Figure 6.2.

The cluster each day belongs to is a discrete value so two classification algo-

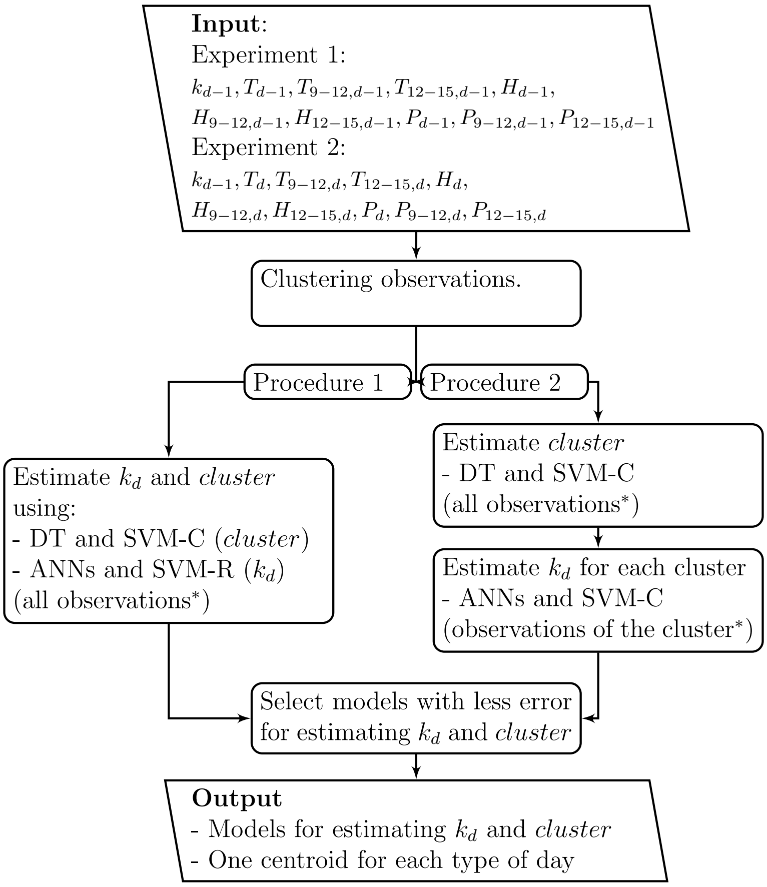

Input:
Experiment 1:
$k_{d-1}, T_{d-1}, T_{9-12,d-1}, T_{12-15,d-1}, H_{d-1},$
$H_{9-12,d-1}, H_{12-15,d-1}, P_{d-1}, P_{9-12,d-1}, P_{12-15,d-1}$
Experiment 2:
$k_{d-1}, T_d, T_{9-12,d}, T_{12-15,d}, H_d,$
$H_{9-12,d}, H_{12-15,d}, P_d, P_{9-12,d}, P_{12-15,d}$
Clustering observations.
Procedure 1
Procedure 2
Estimate k_d and *cluster* using:
- DT and SVM-C (*cluster*)
- ANNs and SVM-R (k_d)
(all observations*)
Estimate *cluster*
- DT and SVM-C
(all observations*)
Estimate k_d for each cluster
- ANNs and SVM-C
(observations of the cluster*)
Select models with less error for estimating k_d and *cluster*
Output
- Models for estimating k_d and *cluster*
- One centroid for each type of day

Figure 6.2: Scheme of the proposed procedures to obtain models to characterize hourly global solar radiation. In procedure 1, only one model is selected to estimate k_d, regardless of cluster. In procedure 2, one model for each cluster is selected to estimate k_d. *training set

rithms are used to predict it: Decision Trees (DT) and SVM-C (Support Vector Machines for Classification). A continuous value predictor method should be used to predict k_d values and, in this case, ANN (Artifical Neural Network) and SVM-R (Support Vector Machine for regression) are used. As two methods are used to predict each of the two dependent variables (k_d and *cluster*), all the combinations of these methods are tested, so there are $2 \times 2 = 4$ methods combinations.

The four method combinations have been tested:

- DT + ANN
- DT + SVM-R

- SVM-C + ANN
- SVM-C + SVM-R

All these models are estimated using a data subset, usually known as training data. The models obtained using training data are checked using test data to estimate the error of each combination. Training data set has 80% of samples while the test data set has the remaining 20% samples. The samples of each set are chosen randomly. Used data set ranges from October 2010 to December 2013.

MATLAB 2014b implementations are used for the ANN and DT methods, while LibSVM library from Chih-Chung Chang and Chih-Jen Lin are used for both SVM-C and SVM-R methods (Chang and Lin, 2011).

The ANN predictor is configured with input, hidden and output layers, and three neurons in the hidden layer. The hyperbolic tangent sigmoid transfer function is used in the input and hidden layer, while linear transfer function is used in the output layer. For the DT method, the MATLAB *fitctree* function is used with default options. SVM-C and SVM-R used here are implemented in an external library, which allows regression and multiple class classification with SVM to be performed. For the SVM-C method the paremeter *nu* has been set to 0.01 while for the SVM-R method the parameter γ for the kernel function has been set to 0.001. These values have been obtained after checking several values to obtain good results, as they are data dependent.

6.2.3 Hourly solar radiation forecasting

Forecasting the hourly solar radiation values for the next day is performed in two phases that allow the next-day clearness index and the next-day type (cluster) to be determined; the hourly profiles of solar radiation are estimated using these two predicted parameters. The process for modeling hourly values of global solar radiation for a day is similar (note the difference between modeling and forecasting). The only difference is in the input data used; the values of the different meteorological parameters of the previous day and predictions of these parameters are used in forecasting while the values of these parameters for the current day are used in modeling process.

Once the models for forecasting daily clearness index and cluster for next day have been built, the process for forecasting hourly values of global solar radiation for next-day is shown in Figure 6.3. First, the cluster and the daily clearness index, k_d, are predicted using the selected models and the proposed independent variables. Then, the centroid of the predicted cluster (hourly values of index $k^*_{h,d}$) is selected and the predicted value of k_d is added to each of the eight hourly values (the centroid) of $k^*_{h,d}$ to obtain the values of hourly clearness index, $k_{h,d}$. Finally, the hourly global horizontal solar radiation is calculated using Eq. 4.17 and Eq. 4.6.

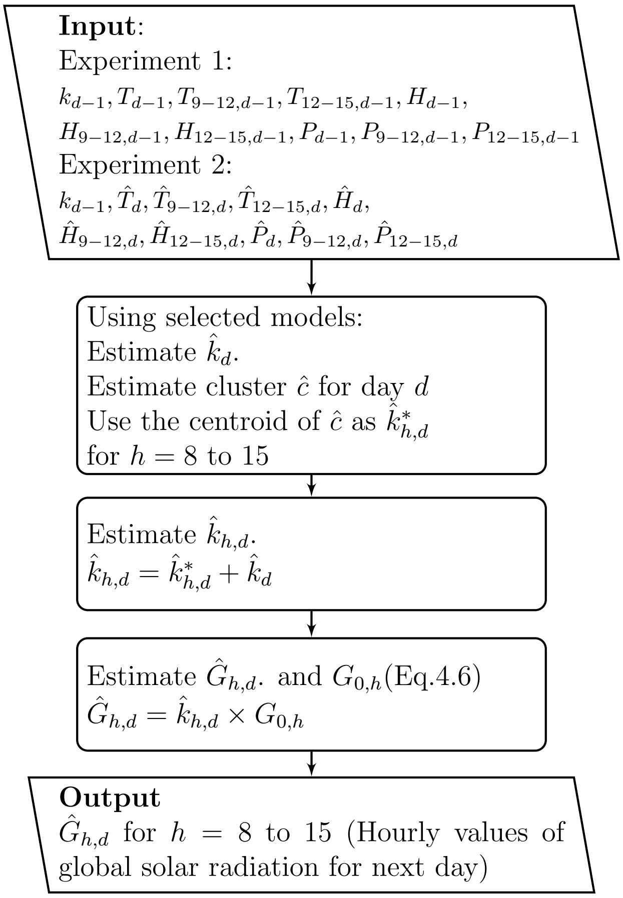

Input:
Experiment 1:
$k_{d-1}, T_{d-1}, T_{9-12,d-1}, T_{12-15,d-1}, H_{d-1},$
$H_{9-12,d-1}, H_{12-15,d-1}, P_{d-1}, P_{9-12,d-1}, P_{12-15,d-1}$
Experiment 2:
$k_{d-1}, \hat{T}_d, \hat{T}_{9-12,d}, \hat{T}_{12-15,d}, \hat{H}_d,$
$\hat{H}_{9-12,d}, \hat{H}_{12-15,d}, \hat{P}_d, \hat{P}_{9-12,d}, \hat{P}_{12-15,d}$
Using selected models:
Estimate $\hat{k}_d$.
Estimate cluster $\hat{c}$ for day d
Use the centroid of $\hat{c}$ as $\hat{k}^*_{h,d}$
for $h = 8$ to 15
Estimate $\hat{k}_{h,d}$.
$\hat{k}_{h,d} = \hat{k}^*_{h,d} + \hat{k}_d$
Estimate $\hat{G}_{h,d}$. and $G_{0,h}$(Eq.4.6)
$\hat{G}_{h,d} = \hat{k}_{h,d} \times G_{0,h}$
Output
$\hat{G}_{h,d}$ for $h = 8$ to 15 (Hourly values of global solar radiation for next day)

Figure 6.3: Proposed procedure for forecasting next-day values of hourly solar radiation.

6.3 Model and forecast evaluations

6.3.1 Experiments

Two experiments have been performed using different input data sets depending on the used independent variables. In the first one, the values of the meteorological variables for the previous day are the independent variables of the analyzed models (forecasting). In the second one, the values of these variables for the same day to be forecasted are the independent variables except for the value of daily clearness index, where the value used corresponds to the previous day (modeling). The second experiment serves as test for using the forecasted meteorological data for the same day provided by a forecasting service.

Table 6.1 summarizes the data used in each experiment (columns 1 and 2) including the input variables used for checking the models built in Experiment 2. The predicted data were collected from the Spanish Weather Service (AEMET) from November to December of 2015. In all cases, the dependent variables are the *cluster* and the daily clearness index.

Independent variables		Forecasting input variables (Exper.2)
Experiment 1	Experiment 2	
k_{d-1}	k_{d-1}	k_{d-1}
$T_{d-1}, T_{9-12,d-1}, T_{12-15,d-1}$	$T_d, T_{9-12,d}, T_{12-15,d}$	$\hat{T}_d, \hat{T}_{9-12,d}, \hat{T}_{12-15,d}$
$H_{d-1}, H_{9-12,d-1}, H_{12-15,d-1}$	$H_d, H_{9-12,d}, H_{12-15,d}$	$\hat{H}_d, \hat{H}_{9-12,d}, \hat{H}_{12-15,d}$
$P_{d-1}, P_{9-12,d-1}, P_{12-15,d-1}$	$P_d, P_{9-12,d}, P_{12-15,d}$	$\hat{P}_d, \hat{P}_{9-12,d}, \hat{P}_{12-15,d}$

Table 6.1: Independent variables used to build the models (cols 1 and 2) and forecasting input variables. Note: $d-1$ refers to the previous day and d current day, $12-15$ refers to the mean value estimated using values from 12 to 15 hours, and $9-12$ is for the mean value estimated using values from 9 to 12 hours, $\hat{X}$ refers to the forecasted value of X, T is temperature, H is humidity and P atmospheric pressure.

6.3.2 Error metrics

Error metrics used here were described in 3.4. *MSE*, *rMSE*, *MAE*, *rMAE* and *Forecast Skill* are used to compare the different proposed models. The persistence model assumes that the conditions at the time of the forecast will not change, that is, the forecast for tomorrow are the values of today. The value of s is a measure of improvement over the persistence (higher values indicate better forecast skill, see 4.5.2). The forecast skill is independent of the specific meteorological or climatologic characteristics of the site under consideration (Coimbra and Kleissl, 2013).

6.3.3 Evaluation of clustering

The first step is to cluster the observations using K-means to obtain four types of daily profiles. Each observation consists of 8 hourly values of variable $k^*_{h,d}$ (Eq.6.1) estimated from the global solar radiation for each day. Table 6.2 shows the amount of elements (in percentages) in each cluster obtained with K-means and the percentage of total radiation received in the days included in the cluster.

Cluster	% Elements in cluster	% Total radiation received in days included in cluster
1	66	72
2	14	11
3	12	10
4	8	7

Table 6.2: Number of observations in each cluster and percentage of radiation received in days included in cluster

Clusters centroids are shown in Figure 6.4. Cluster 1 groups clear days, cluster 2 represents days with overcast variable among the day, cluster 3 represents days that are mainly overcast in the morning and cluster 4 represents days that are mainly overcast in the afternoon. Figure 6.5 shows the observations in each cluster. As it can be observed, most of the days are similar to the centroid of the cluster they belong to. These results confirm the hypothesis that it is possible to group hourly global radiation profiles throughout the day using 4 different classes.

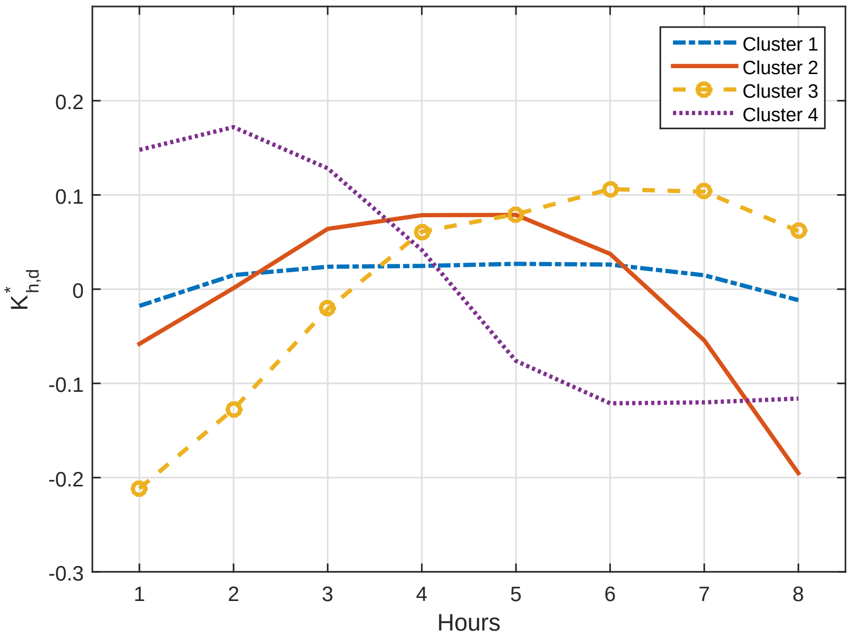


Figure 6.4: Values of index $k^*_{h,d}$ (centroids) for each cluster.

6.3.4 Evaluation of models

The second step in the data analysis is to determine the relationships between the daily clearness index and the meteorological parameters using two data mining techniques. This analysis is performed using two different procedures as explained in Section 6.2.2 and using two different input data sets (see Table 6.1). In the first procedure, the models to obtain clusters and daily clearness index are fitted separately and all training set observations are used to fit these models. In the second procedure, once the models to obtain the cluster are fitted and the clusters built, different models are fitted for the observations in each cluster to obtain the daily clearness index.

Each procedure has been used for two different data sets. The first set of independent variables consists of variables of the previous day (Experiment 1). The second one corresponds to variables of the day to forecast or to model (Experiment 2), except for the daily clearness index, where the value of the previous day is used.

The cluster estimating results for the test data set, when DT is used, are that 59.6% of data are correctly classified while the 65.6% of data are correctly classified

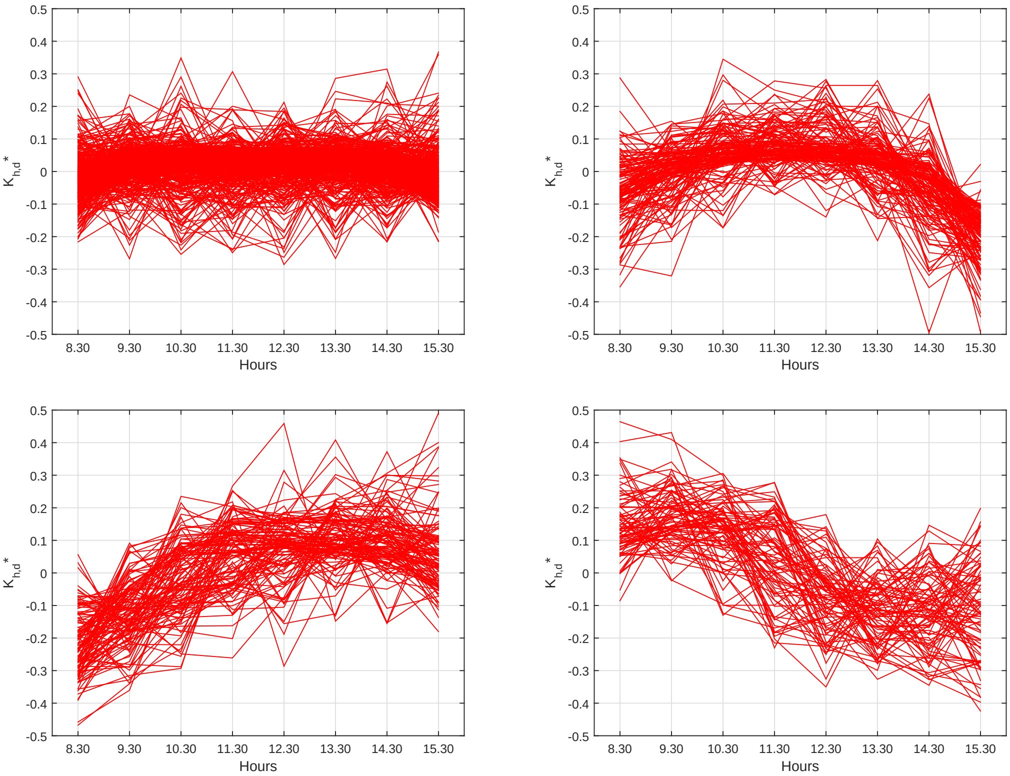


Figure 6.5: Values of index $k^*_{h,d}$ for the days included in each cluster. Each line corresponds to the values of this index for a day.

when SVM-C is used. A joint assessment of the models to estimate the cluster and the models to estimate daily clearness index has been performed.

The proposed models have been evaluated using the MAE, $rMAE$ and $RMSE$ for hourly global solar radiation. First, models are evaluated using training set. Errors estimated for each experiment for training set are shown in Table 6.3.

The combination of SVM-C+SVM-R in which a different SVM-R model for each cluster to estimate the daily clearness index is used, allows to obtain lower errors for both data inputs (Experiments 1 and 2). The value of $rMAE$ for Experiment 1 is 13.3% while for Experiment 2 is 10.5%. The $RMSE$ values are 22.2% and 16.8% for Experiment 1 and 2 respectively. All these values were obtained for training sets.

		Experiment 1		Experiment 2	
Method		$MAE(rMAE)$	$RMSE$	$MAE(rMAE)$	$RMSE$
cluster	k_d	$W/m^2(\%)$	$W/m^2(\%)$	$W/m^2(\%)$	$W/m^2(\%)$
DT	ANN$^{(1)}$	102 (19.0)	143 (26.7)	71 (13.3)	103 (19.3)
DT	SVM-R$^{(1)}$	94 (17.2)	139 (25.4)	68 (12.5)	102 (18.9)
SVM-C	ANN$^{(1)}$	97 (18.2)	137 (25.6)	68 (12.6)	98 (18.3)
SVM-C	SVM-R$^{(1)}$	90 (16.4)	133 (24.3)	65 (11.9)	97 (17.9)
DT	ANN$^{(2)}$	92 (17.5)	135 (25.6)	67 (12.6)	100 (18.8)
DT	SVM-R$^{(2)}$	79 (14.3)	128 (23.3)	61 (11.3)	96 (17.8)
SVM-C	ANN$^{(2)}$	87 (16.5)	129 (24.3)	63 (11.8)	94 (17.7)
SVM-C	SVM-R$^{(2)}$	**73 (13.3)**	**122 (22.2)**	**57 (10.5)**	**91 (16.8)**

Table 6.3: Training set: errors for each analyzed combination of methods for Experiment 1 (values of meteorological variables for the previous day) and Experiment 2 (values of meteorological variables for the same day to be modeled are the independent variables except the value of daily clearness index). $^{(1)}$ The same model is fitted for all observations (Procedure 1). $^{(2)}$ One model is fitted for the observations of each cluster (Procedure 2).

6.3.5 Evaluation of forecasts

Test sets were used to evaluate the forecasts of different models. Models obtained in Experiment 1 were evaluated with the test set obtained from the measurement not used to build the models while models obtained in Experiment 2 need to be evaluated using forecasted data of the meteorological variables instead of the measured data. The data used in this case are the prediction obtained from the AEMET (Spanish Weather Service) from November to December 2015. Estimated errors for each experiment for test sets are shown in Table 6.4.

The lowest $rMAE$ is 16.7% in the case of Experiment 1 while it is 15.2% in the case of Experiment 2. These results are obtained when SVM-C and SVM-R are used for estimating the cluster and the daily clearness index respectively and when a different model is estimated for each cluster.

Tables 6.5 and 6.6 show detailed MAE, $rMAE$ and $RMSE$ values for each experiment, procedure, cluster and data set (training and test). Not in all cases the SVM-C+SVM-R is the best model for each cluster, but is the best model taking into account all clusters.

Method		Experiment 1		Experiment 2	
		$MAE(rMAE)$	$RMSE$	$MAE(rMAE)$	$RMSE$
cluster	k_d	$W/m^2(\%)$	$W/m^2(\%)$	$W/m^2(\%)$	$W/m^2(\%)$
DT	ANN(1)	107 (19.9)	150 (27.9)	126 (24.3)	170 (32.7)
DT	SVM-R(1)	101 (18.3)	148 (26.9)	111 (21.3)	153 (29.4)
SVM-C	ANN(1)	104 (19.3)	145 (27.0)	121 (23.3)	159 (30.6)
SVM-C	SVM-R(1)	97 (17.6)	144 (26.0)	103 (19.8)	141 (27.1)
DT	ANN(2)	112 (21.1)	161 (30.1)	120 (23.1)	180 (34.6)
DT	SVM-R(2)	107 (19.3)	163 (29.3)	106 (20.3)	160 (30.8)
SVM-C	ANN(2)	102 (18.5)	150 (27.2)	105 (20.2)	152 (29.2)
SVM-C	SVM-R(2)	**97 (16.7)**	**147 (25.3)**	**79 (15.2)**	**119 (22.9)**

Table 6.4: Test sets: errors for each analyzed combination of methods for Experiment 1 (values of meteorological variables for the previous day) and Experiment 2 (forecasts of meteorological variables for the same day to be modeled are the independent variables except the value of daily clearness index). (1) The same model is fitted for all observations (Procedure 1). (2) One model is fitted for the observations of each cluster (Procedure 2).

Method		$MAE(rMAE)$				$RMSE$			
cluster	k_d	$W/m^2(\%)$				$W/m^2(\%)$			
Cluster		1	2	3	4	1	2	3	4
		Experiment 1							
DT	ANN(1)	94 (0.17)	83 (0.19)	125 (0.25)	158 (0.31)	134 (0.22)	110 (0.22)	168 (0.34)	204 (0.41)
DT	SVM-R(1)	84 (0.14)	76 (0.17)	125 (0.25)	163 (0.31)	127 (0.21)	103 (0.21)	169 (0.34)	210 (0.42)
SVM-C	ANN(1)	93 (0.16)	77 (0.18)	116 (0.23)	138 (0.27)	133 (0.22)	103 (0.21)	158 (0.32)	180 (0.36)
SVM-C	SVM-R(1)	**82 (0.14)**	**70 (0.16)**	**116 (0.23)**	**144 (0.27)**	**126 (0.21)**	**94 (0.19)**	**159 (0.32)**	**187 (0.38)**
DT	ANN(2)	84 (0.14)	77 (0.18)	122 (0.29)	139 (0.33)	130 (0.21)	105 (0.21)	155 (0.32)	177 (0.35)
DT	SVM-R(2)	70 (0.11)	67 (0.15)	110 (0.26)	124 (0.28)	128 (0.21)	94 (0.19)	141 (0.29)	158 (0.32)
SVM-C	ANN(2)	82 (0.14)	74 (0.17)	110 (0.26)	119 (0.29)	127 (0.21)	99 (0.20)	144 (0.29)	154 (0.31)
SVM-C	SVM-R(2)	**68 (0.11)**	**61 (0.14)**	**97 (0.23)**	**101 (0.23)**	**126 (0.21)**	**86 (0.17)**	**128 (0.26)**	**130 (0.26)**
		Experiment 2							
DT	ANN(1)	61 (0.11)	67 (0.15)	91 (0.20)	128 (0.26)	91 (0.15)	93 (0.19)	123 (0.25)	164 (0.33)
DT	SVM-R(1)	56 (0.10)	67 (0.15)	91 (0.20)	130 (0.26)	88 (0.14)	93 (0.19)	125 (0.25)	167 (0.34)
SVM-C	ANN(1)	60 (0.10)	63 (0.14)	83 (0.18)	115 (0.24)	88 (0.15)	87 (0.18)	114 (0.23)	148 (0.30)
SVM-C	SVM-R(1)	**55 (0.09)**	**63 (0.14)**	**84 (0.18)**	**118 (0.24)**	**86 (0.14)**	**87 (0.18)**	**117 (0.24)**	**151 (0.30)**
DT	ANN(2)	55 (0.09)	73 (0.17)	94 (0.22)	119 (0.27)	87 (0.14)	98 (0.20)	123 (0.25)	150 (0.30)
DT	SVM-R(2)	50 (0.08)	64 (0.15)	88 (0.20)	112 (0.26)	87 (0.14)	89 (0.18)	115 (0.23)	142 (0.29)
SVM-C	ANN(2)	54 (0.09)	65 (0.16)	83 (0.19)	104 (0.24)	85 (0.14)	88 (0.18)	111 (0.23)	135 (0.27)
SVM-C	SVM-R(2)	**48 (0.08)**	**59 (0.14)**	**78 (0.180)**	**94 (0.22)**	**85 (0.14)**	**83 (0.17)**	**104 (0.21)**	**124 (0.25)**

Table 6.5: Training set: errors by cluster for each analyzed combination of methods for Experiment 1 (values of meteorological variables for the previous day) and Experiment 2 (values of meteorological variables for the same day to be modeled are the independent variables except the value of daily clearness index). (1) The same model is fitted for all observations (Procedure 1). (2) One model is fitted for the observations of each cluster (Procedure 2).

Method		$MAE(rMAE)$				$RMSE$			
cluster	k_d	$W/m^2(\%)$				$W/m^2(\%)$			
Cluster		1	2	3	4	1	2	3	4
		Experiment 1							
DT	ANN$^{(1)}$	98 (0.17)	91 (0.21)	136 (0.28)	167 (0.32)	140 (0.23)	120 (0.24)	179 (0.37)	211 (0.42)
DT	SVM-R$^{(1)}$	89 (0.15)	86 (0.19)	139 (0.28)	174 (0.32)	135 (0.22)	117 (0.24)	185 (0.38)	221 (0.44)
SVM-C	ANN$^{(1)}$	94 (0.17)	90 (0.20)	133 (0.27)	165 (0.32)	135 (0.22)	117 (0.24)	174 (0.36)	208 (0.42)
SVM-C	SVM-R$^{(1)}$	**85 (0.15)**	**85 (0.19)**	**136 (0.27)**	**171 (0.32)**	**129 (0.21)**	**113 (0.23)**	**180 (0.37)**	**219 (0.44)**
DT	ANN$^{(2)}$	94 (0.16)	122 (0.26)	159 (0.38)	193 (0.50)	142 (0.23)	163 (0.33)	196 (0.40)	236 (0.47)
DT	SVM-R$^{(2)}$	88 (0.14)	123 (0.26)	152 (0.35)	187 (0.47)	148 (0.24)	168 (0.34)	189 (0.39)	227 (0.46)
SVM-C	ANN$^{(2)}$	100 (0.18)	107 (0.23)	170 (0.33)	80 (0.25)	148 (0.24)	132 (0.27)	204 (0.42)	103 (0.21)
SVM-C	SVM-R$^{(2)}$	**96 (0.16)**	**105 (0.23)**	**165 (0.32)**	**74 (0.25)**	**156 (0.26)**	**129 (0.26)**	**202 (0.41)**	**90 (0.18)**
		Experiment 2							
DT	ANN$^{(1)}$	68 (0.12)	77 (0.17)	111 (0.25)	149 (0.30)	101 (0.17)	104 (0.21)	148 (0.30)	191 (0.38)
DT	SVM-R$^{(1)}$	63 (0.11)	76 (0.17)	107 (0.24)	150 (0.29)	97 (0.16)	105 (0.21)	145 (0.30)	195 (0.39)
SVM-C	ANN$^{(1)}$	63 (0.11)	76 (0.17)	113 (0.25)	151 (0.30)	93 (0.15)	103 (0.21)	148 (0.30)	192 (0.39)
SVM-C	SVM-R$^{(1)}$	**59 (0.10)**	**75 (0.17)**	**110 (0.24)**	**151 (.030)**	**89 (0.15)**	**103 (0.21)**	**146 (0.30)**	**196 (0.39)**
DT	ANN$^{(2)}$	66 (0.11)	97 (0.22)	120 (0.30)	162 (0.41)	104 (0.17)	132 (0.27)	154 (0.32)	202 (0.41)
DT	SVM-R$^{(2)}$	62 (0.11)	92 (0.20)	113 (0.28)	149 (0.37)	104 (0.17)	129 (0.26)	146 (0.30)	187 (0.37)
SVM-C	ANN$^{(2)}$	74 (0.14)	79 (0.18)	115 (0.24)	95 (0.24)	114 (0.19)	106 (0.21)	145 (0.30)	114 (0.23)
SVM-C	SVM-R$^{(2)}$	**71 (0.13)**	**73 (0.16)**	**107 (0.22)**	**92 (0.24)**	**113 (0.19)**	**95 (0.19)**	**137 (0.28)**	**112 (0.22)**

Table 6.6: Test sets: errors by cluster for each analyzed combination of methods for Experiment 1 (values of meteorological variables for the previous day) and Experiment 2 (forecasts of meteorological variables for the same day to be modeled are the independent variables except the value of daily clearness index). $^{(1)}$ The same model is fitted for all observations (Procedure 1). $^{(2)}$ One model is fitted for the observations of each cluster (Procedure 2).

Locations	$rMAE$(%)	$RMSE$(%)	s(%)
Southern Spain (Lorenz et al., 2009)	12.2-20.4	20.8-37.1	2-36
persistence	16.6	32.1	
Southern Spain (Perez et al., 2013)	13-21	22-29	17-37
persistence	19	35	
Rome (Pierro et al., 2015)	16.3-22.3	26.4-34.1	26-29
persistence	25	47	
SVM-C + SVM-R (Experiment 1)	16.7	25.3	23.0
persistence	25.8	32.7	
SVM-C + SVM-R (Experiment 2)	15.2	22.9	43.9
persistence	24.0	40.8	

Table 6.7: Comparison of the performance of the SVM-C+SVM-R models.

Finally, to compare the results obtained for the model that gives the best results for both experiments with previously proposed models, the forecast skills over 24 hour persistence forecasts (s) were also estimated. Table 6.7 compares the benchmark accuracy range previously found in (Lorenz et al., 2009) and (Perez et al., 2013) for the southern Spain and (Pierro et al., 2015) for Rome (Italy) with the performance obtained by proposed model.

Regarding the results obtained in Experiment 1 the proposed model gives values of $rMAE$ and $RMSE$ inside the benchmark of the values obtained for Spanish locations and similar to those obtained for Rome, except in this case for $RMSE$ that is slightly lower. The value of forecast skill over 24 hour persistence obtained is 23% and is also inside the benchmark of Spanish locations but again lower that values for Rome. It should be noted that the $RMSE$ of the persistence model for this location is greater than the value obtained for Spanish locations. This suggest that the data used for Spanish locations correspond to a period with more sunny days, as pointed out by authors.

The errors obtained in Experiment 2 are also inside the benchmark of the values obtained for Spanish locations, $rMAE$ is 15.2% and $RMSE$ is 22.9%. The value of parameter s is 43.9%, greater than the previosly reported. This could be due to the fact that only data from November and December were used where the number of sunny days is lower and the performance of persistence model is worse.

6.4 Conclusions

A new approach based on the use of different data mining techniques to model and to improve the next-day prediction of hourly global solar radiation has been developed. Modeling and prediction are conducted in two different phases and using different data mining techniques in each one. A clustering algorithm is used to identify how many different day types there are in the first phase, and in the second phase, different classification algorithms are combined with regression algorithms to obtain the parameters to forecast hourly global solar radiation with four method combinations: decision trees, artificial neural networks, support vector machine for regression and support vector machine for classification. The classification algorithms are used to estimate the cluster to which the day belongs to. Regression algorithms are used to estimate the daily clearness index. These two estimated parameters and the proposed centroids are used to predict the hourly global solar radiation. Four different types of daily profiles have been estimated. These profiles correspond to clear (cloudless), overcast (variable among the day), overcast mainly in the morning and overcast mainly in the afternoon.

Two procedures have been proposed and two experiments have been performed using different input data sets depending on the used independent variables. Results show that the best method combination is SVM-C to estimate the cluster of each observation and SVM-R to estimate the daily clearness index.

Regarding the results obtained in Experiment 1, when the independent variables are the values of the meteorological parameters for the previous day, the $rMAE$ value for the best model is 16.7% and the $RMSE$ value 23.5%. The results obtained for $rMAE$ and $RMSE$ are 15.2% and 22.9% respectively for Experiment 2, when the independent variables are the forecasts of meteorological variables for the same day to be forecasted except for the value of daily clearness index that corresponds to the previous day. These values are inside the benchmark previously proposed by several authors. The estimated value of forecast skill over 24 hours persistence is 43.9% that is greater than the values previously reported; this may be due to the period of data that has been used to evaluate the model predictions as only data for November and December 2015 were used.

Bibliography

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27, software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Coimbra, C., Kleissl, J., 2013. Solar Resource Assessment and Forecasting. Elsevier, Waltham, Massachusetts, Ch. Chapter 8: Overview of Solar-Forecasting Methods and Metric for Accuracy Evaluation, pp. 171–194.

Lorenz, E., Remund, J., Müller, S. C., Traunmüller, W., Steinmaurer, G., Pozo, D., Ruiz-Arias, J. A., Fanego, V. L., Ramirez, L., Romeo, M. G., et al., 2009. Benchmarking of different approaches to forecast solar irradiance. In: 24th European photovoltaic solar energy conference, Hamburg, Germany. Vol. 21. p. 25.

Perez, R., Lorenz, E., Pelland, S., Beauharnois, M., Knowe, G. V., Jr., K. H., Heinemann, D., Remund, J., Müller, S. C., Traunmüller, W., Steinmauer, G., Pozo, D., Ruiz-Arias, J. A., Lara-Fanego, V., Ramirez-Santigosa, L., Gaston-Romero, M., Pomares, L. M., 2013. Comparison of numerical weather prediction solar irradiance forecasts in the us, canada and europe. Solar Energy 94, 305 – 326.
URL `http://www.sciencedirect.com/science/article/pii/S0038092X13001886`

Pierro, M., Bucci, F., Cornaro, C., Maggioni, E., Perotto, A., Pravettoni, M., Spada, F., 2015. Model output statistics cascade to improve day ahead solar irradiance forecast. Solar Energy 117, 99 – 113.
URL `http://www.sciencedirect.com/science/article/pii/S0038092X15002212`

Chapter 7

Forecasting and assessing the performance of photovoltaic facilities

To ensure optimal performance of small and medium power solar photovoltaic plants, it is essential to monitor and evaluate the parameters of these plants. However, there are not standard solutions in the market to allow monitoring and remote evaluation based on the prediction of expected production of energy. In this chapter, a software architecture that allows monitoring and evaluation using forecasted values of solar radiation is presented. As an example of the use of the proposed remote evaluation method, it is presented an application which aims to monitor and supervise 40 grid-connected photovoltaic systems with different configurations and subsystems.

The solar photovoltaic plants connected to the grid have experienced a strong development in recent years (Ciesielska et al., 2011). However, the future development of photovoltaics in urban environments depends on small systems that can be installed taking advantage of sunny areas on terraces and roofs of existing buildings, and on integrating photovoltaic panels as construction elements in designing new architectural constructions.

All photovoltaic systems are designed to produce the maximum possible energy, maximizing the investment benefits. Supervising the operation of a photovoltaic system is not trivial and it is further complicated by the lack of a standard for communications between main elements of the plant and data acquisition systems.

For large photovoltaic systems, monitoring tools are designed *key on hand* to provide analysis of information of the plant in real time and by remote access to data. However, for small systems, the order of 100 kWp or less, these systems should be able to implement reliable monitoring tools, simple and low cost, so they can be easily made within the cost of the project and under the hypothesis that these plants do not have specialized staff for monitoring and control. Inverter manufacturers make available to customers applications that provide an overview of the system status and energy production. They incorporate internally or externally data acquisition systems that allow collecting meteorological variables necessary for the evaluation of the system. Often, these applications are closed, incompatible with inverters from other manufacturers and, in some cases, they only allow downloading stored data, leaving the data evaluation to be performed subsequently by the customer, who does not have enough experience in this field.

A variety of tools have been proposed for monitoring photovoltaic plants: (Figueiredo and da Costa, 2008) propose a system controller PLC to control the production and consumption of energy (project developed in an experimental park) of each plant, interconnected by a network and managed by a SCADA system. (Kalaitzakis et al., 2003) proposes a system in which data acquisition systems are connected wirelessly with data collection systems and the data is sent to a central server which makes the data accessible to any user on the Internet. (Wang and Liu, 2007) proposes a set of PLC controllers for real-time readings of several energy systems (wind turbine converter, inverter and battery), connected to a intranet and also includes control of wind turbines and a web interface for user. (Guozhen et al., 2009) proposes improved SCADA systems with two important aspects: security (through authentication, data encryption and user roles) and reliability (using redundant systems). The problem with these proposals is that, on one hand, some of these systems involve the use of additional hardware such as the PLC's and local computers to store data and others propose wireless connections, which usually cause problems in sites without maintenance personnel.

In this chapter an evaluating system model for solar photovoltaic power plants is presented. This system has been integrated into the monitoring system proposed in (Martínez Marchena I, 2014). The proposed system allows us including the evaluation in the developed framework.

7.1 Photovoltaic Monitoring System

In a photovoltaic system electrical parameters are recorded in the inverter. The recorded parameters are at least at input and output of the inverter power, voltage and current both at input and output of the inverter. In some plants, are also recorded the incident irradiance on the modules plane and the module temperature, but unfortunatelly this is not the case for most of small photovoltaic installation. All measurements recorded at each plant are instantaneous and are recorded at a frequency that is defined for each plant. From these parameters, measures are defined in different time scales for performance evaluation of the plants:

- hourly: they represent the accumulated or mean values over a hour depending on the parameter. Values are calculated from the instantaneous values.
- daily: they represent the accumulated or mean values over a day. Values can be calculated both from the instantaneous values and from the hourly values.
- global: representing accumulated values over the life of the plant.

7.1.1 System description

In figure 7.1 the system architecture for monitoring and evaluating solar photovoltaic plants is presented, which is described in (Martínez Marchena I, 2014). Figure 7.1 summarizes the various configurations that can be integrated into monitoring systems that are developed from this architecture, for example, it is possible to integrate facilities that have physical monitoring systems (dataloggers) as well as facilities where data are obtained directly from inverters. The data reading from plants is done through software elements called OPC servers, which are developed for each technology (inverter, datalogger, etc.) and through a client software (which plays the role of OPC client) data is stored in a central database, including the structural information and monitoring data collected from facilities over time.

The system uses OPC technology (Holley, 2004) to read data from electrical equipment, this way the interface that data providers (OPC servers) present to the client applications that implement OPC clients is standard. Also, OPC technology

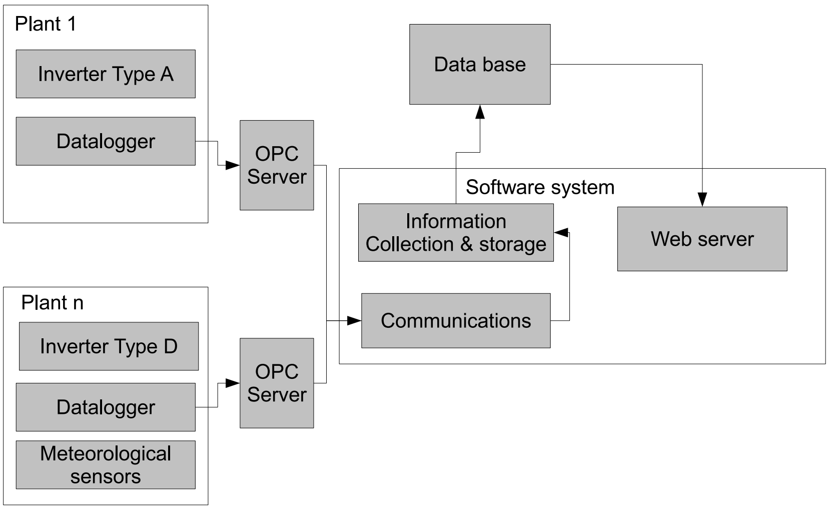


Figure 7.1: Main system architecture

enables the existence of applications with different functionality using the same OPC servers: synchronize data, display data from OPC servers, evaluating data, etc. Scheduled tasks are used to read and evaluate data from plants running on a machine and as a result the users (plant owners, staff responsible for maintenance, etc.) receive e-mail messages and SMS messages that keep them informed of the status of their facilities, depending on the level of access that each one is assigned.

7.2 Assessing photovoltaic facilities

Assessment models can be used to supervise the performance of the solar power plants. As proposed in (Martínez Marchena I, 2014), evaluation parameters can be implemented easily as performance indexes: daily final yield and daily enegy balances (daily energy produced and daily energy forecasted).

The daily final yield, $Y_{f,day}$, is defined as the useful output energy of the PV per kW_p installed:

$$Y_{f,day} = \frac{E_{day}}{P_{STC}} \tag{7.1}$$

where P_{STC} is the nominal power of the installed photovoltaic array at standard test conditions (1 kM/m^2 of solar irradiance and 25°cell temperature).

The useful daily output energy or daily energy supplied by a solar power plant is defined as:

$$E_{day} = \int_{day} P_{AC}(t)d(t) \approx \sum_{j=1}^{n} P_{AC}^{j} \Delta t \tag{7.2}$$

where n is the number of measurements along the day and P_{AC}^{j} are the recorded values of the power generated at the inverter output.

The proposal is to assess photovoltaic facilities in two phases. In the first phase, the evaluation is done by comparing the values obtained for the daily final yield and the useful daily output energy when they are estimated using two different methods. On the one hand, these values are estimated directly using the recorded values in the inverter. In the other hand, these parameters are estimated using a model for the performance of photovoltaic plant and the forecasted values of hourly global radiation. In this phase only daily values are compared. Taking into account the results of this comparison, a second phase could be activated in order to compare hourly values.

7.2.1 Models for estimating the performance of a PV system

As it has been reported previously, see for instance Ayompe et al. (2011), the inverter power output has a linear relationship with solar irradiance if it is not considered the effect of the temperature. This fact can be observed in Fig. 7.2.

We propose to include this effect in the model that allows to estimate the hourly energy generated at the output of the inverter (E_h^*) as an extension of the estimation of the power generated (P_{AC}^*), according to the expression:

$$E_h^* = E_{h,STC} \frac{G_{h,\beta}^*}{1000}(1 + \gamma(T_{h,mod} - 25))GL \tag{7.3}$$

where, $G_{h,\beta}^*$ is the forecasted hourly global radiation on the surface of the modules, β is the inclination of the modules, γ is the temperature coefficient of

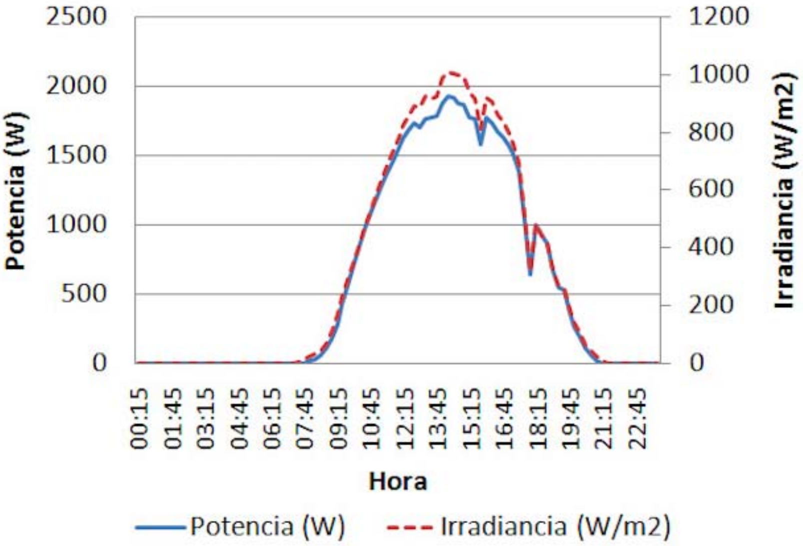


Figure 7.2: Instantaneous values of solar radiation and photovoltaic power

P_m, $T_{h,mod}$ is the mean hourly module temperature and GL is the global losses coefficient of the system. Eq.7.3 is obtained from the expression proposed by Osterwald (1986) considering hourly values. The proposed expression includes not only the losses produced by the temperature but also other losses (soiling, spectral losses and so on).

The predictions of hourly global radiation on the surface of the modules are estimated in the following way:

1. Forecasting the hourly global radiation values on horizontal surface with the model proposed in Chapter 6.2.2 and meteorological variables recorded by weather agencies.

2. Obtaining the values of direct, diffuse and reflected hourly global radiation on horizontal surface.

3. Obtaining the values of global radiation on the surface of the modules.

The results obtained with this model can be used for estimating the daily energy supplied to the grid, E^*_{day}, which is calculated using the expression 7.4:

$$E^{*}_{day} = \int_{day} P^{*}_{AC}(t)d(t) \approx \sum_{j=1}^{n} E_h^{*(j)} \tag{7.4}$$

where n is the number of hours for a day.

The proper operation of plants can be evaluated by comparing the values of E^{*}_{day} with the values of daily energy produced, E_{day}, obtained with Eq.7.2.

Similarly, for detecting problems in plants operation the value of daily final yield, Eq.7.1, has been compared to the estimated daily final yield, $Y^{*}_{f,day}$, calculated with the expression 7.5:

$$Y^{*}_{f,day} = \frac{E^{*}_{day}}{P_{STC}} \tag{7.5}$$

7.2.2 Statistical models for assessing solar plants

With the estimated values described in the previous section and the corresponding measured values, the system is capable of checking the performance of the plants by using the methodology described in (Martínez Marchena I, 2014). The mean values of previously described parameters for photovoltaic solar plants of the same technology (modules and inverters) has been estimated as the standard deviation of them. For evaluating the operation of each plant, a statistical analysis of the differences between the values of estimated parameters and measured parameters is implemented. Thus, for each new recorded or calculated value of some parameter X and its corresponding estimation, X^{*}, at hour i the difference among them is analyzed according to the criteria (significance level 5%):

$$\begin{aligned} d_X^{(i)} &= X^{(i)} - X^{*(i)}; \\ & if\ d_X^{(i)} \notin [-1.96\hat{\sigma}, +1.96\hat{\sigma}]\ then\ mark\ (i), \end{aligned} \tag{7.6}$$

where $\hat{\sigma}$ is the sample standard deviation of X. All the marked values (i) will indicate problems of operation.

For the generated energy, $E_d^{(i)}$, and its corresponding estimation using Eq.7.3,

$E_d^{*(i)}$, the criteria is as follows:

$$\begin{aligned} d_{P_{AC}}^{(i)} &= E_d^{(i)} - E_d^{*(i)}; \\ & if\; d_{P_{AC}}^{(i)} \notin [-1.96\hat{\sigma}, +1.96\hat{\sigma}]\; then\; mark\; (i), \end{aligned} \tag{7.7}$$

For the daily final yield, $Y_{f,day}$, the expression used to decide if there is any operation problem taking into account the expression Eq.7.6 particularized for this parameter is the following:

$$\begin{aligned} d_{Y_{f,day}}^{(i)} &= Y_{f,day}^{(i)} - Y_{f,day}^{*(i)}; \\ & if\; d_{Y_{f,day}}^{(i)} \notin [-1.96\hat{\sigma}, +1.96\hat{\sigma}]\; then\; mark\; (i), \end{aligned} \tag{7.8}$$

Fig. 7.3 shows the algorithm description for the daily assessment of a PV plant. Recorded measures are used as input and estimated measures are calculated from these inputs. Then a test is performed to check if the difference between measured and estimated daily yield is statistically significant. If these values are statistically equals (for a specified significance level) the algorithm ends without detecting problems, while if they are different, each hourly value obtained from recorded power values is compared with the corresponding estimated and if the difference is statistically significant, the corresponding hour is marked. The algorithm ends with operational problems detected and the marked hours are returned.

7.3 Example scenario

From the architecture described in the previous section, monitoring and evaluating systems for plants are developed, integrating in the software of each plant components corresponding to those subsystems which are present. As example of developed systems, this section presents an application which has been developed for the *Agencia Municipal de la Energía del Ayuntamiento de Málaga.* The application allows remote monitoring and evaluation of 40 different solar power facilities, mostly located in public schools and government buildings. All facilities are located in Málaga.

Inverter technologies present in the different plants include Sunways, Ingeteam, Mastervolt, SMA, PowerOne, SolarMax and Fronius. On each power plant floor

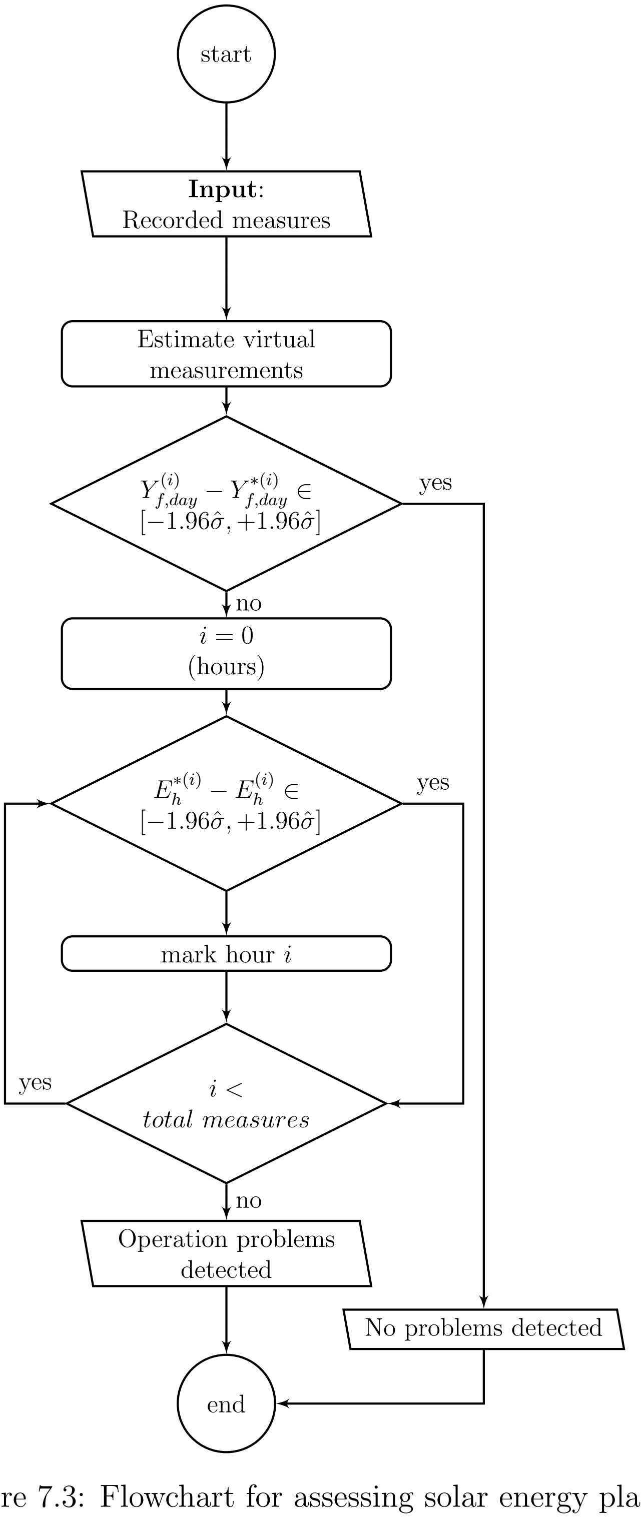

start
Input:
Recorded measures
Estimate virtual measurements
$Y^{(i)}_{f,day} - Y^{*(i)}_{f,day} \in [-1.96\hat{\sigma}, +1.96\hat{\sigma}]$
yes
no
$i = 0$ (hours)
$E^{*(i)}_{h} - E^{(i)}_{h} \in [-1.96\hat{\sigma}, +1.96\hat{\sigma}]$
yes
mark hour i
yes
$i <$ *total measures*
no
Operation problems detected
No problems detected
end

Figure 7.3: Flowchart for assessing solar energy plants.

the existing monitoring system consists of a datalogger that depends on the characteristics of the inverter manufacturer (Meteocontrol, Solarlog, Mastervolt, etc.). With the objective of standardize the acquisition of data, OPC servers have been developed to allow standard access to available data on each technology to integrate them into the system.

The values obtained for serveral days without any detected problems are shown in Fig. 7.4.

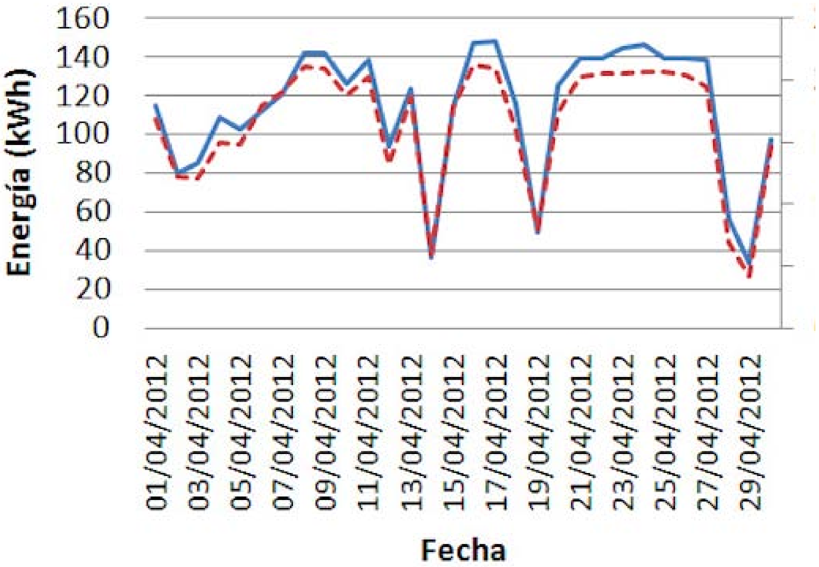


Figure 7.4: Daily parameters

In figure 7.5 and 7.6 operational problems detected with developed software in two different plants are shown.

In the first one, connection problems detected in one of the inverters are shown, while in the second figure, joint analysis of operation of two inverters from the same power plant allows to detect a malfunction of one of them.

7.4 Conclusions

This chapter presents an assessing model for photovoltaic facilities. This model has been integrated in a monitoring framework for solar photovoltaic plants. With

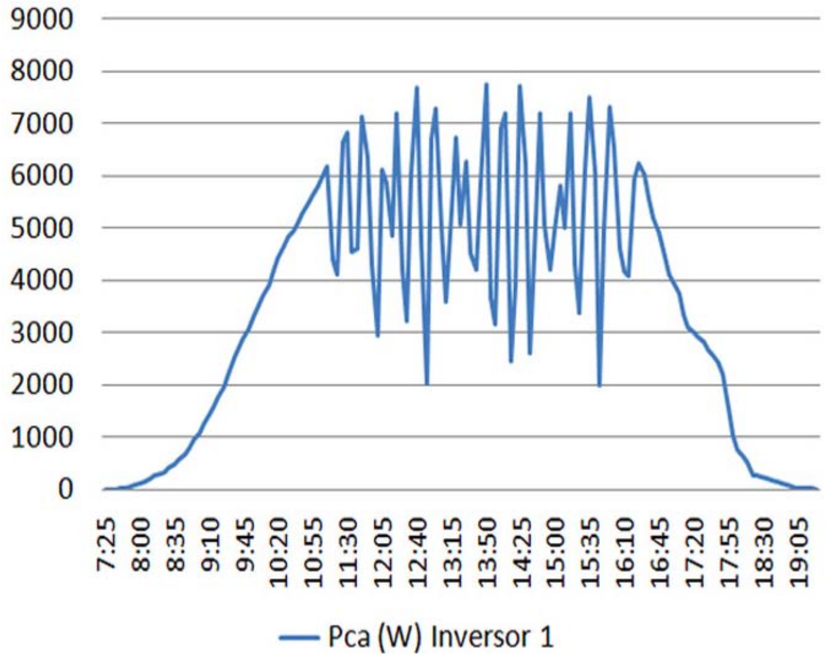


Figure 7.5: Detecting problems in the inverter

the solution proposed applications that integrate a single tool monitoring photovoltaic systems connected to network with inverters or data acquisition systems from different technologies, meter reading and production analysis procedures, evaluation, fault detection and alarm generation of the plant. This proposal enables evaluation of photovoltaic plants remotely using a single program, and without relying on software developed by inverters manufacturers. In addition, this assessment allows rapid action in a plant when malfunctions are detected in the same, thanks to the sending alerts system and remote access to plant data which facilitates maintenance tasks and increases the profitability of photovoltaic systems.

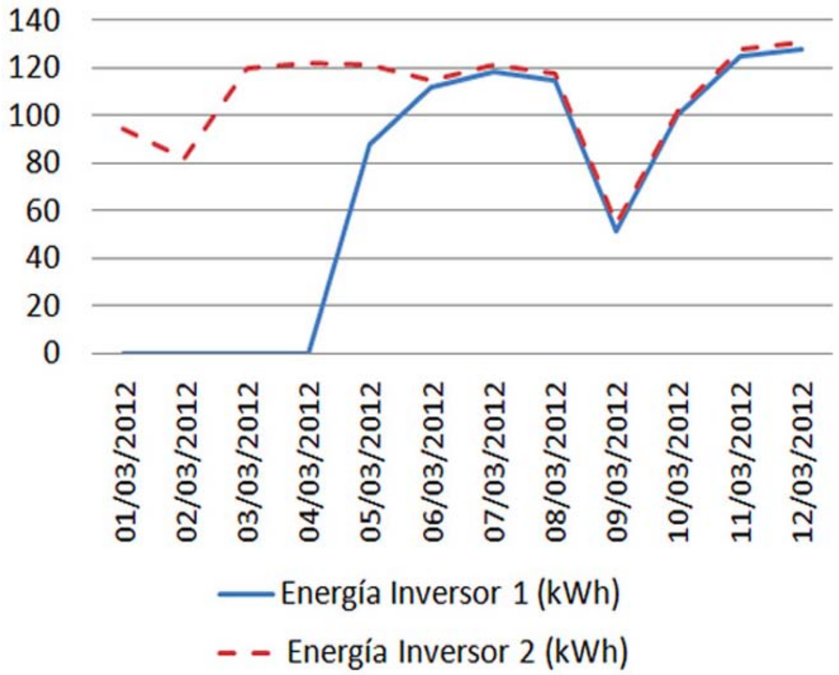


Figure 7.6: Detecting problems in one plant with two inverters

Bibliography

Ayompe, L., Duffy, A., McCormack, S., Conlon, M., 2011. Measured performance of a 1.72 kw rooftop grid connected photovoltaic system in ireland. Energy Conversion and Management 52 (2), 816 – 825.
URL http://www.sciencedirect.com/science/article/pii/S0196890410003730

Ciesielska, J., Concas, G., Despotou, E., Fontaine, B., Garbe, K., Fraile-Montoro, D., et al., 2011. Global market outlook for photovoltaics until 2015. European Photovoltaic Industry Association (EPIA).

Figueiredo, J. M., da Costa, J. M. G. S., May 2008. An efficient system to monitor and control the energy production and consumption. In: 2008 5th International Conference on the European Electricity Market. pp. 1–6.

Guozhen, H., Tao, C., Changsong, C., Shanxu, D., 2009. Solutions for scada system communication reliability in photovoltaic power plants. In: Power Electronics and Motion Control Conference, 2009. IPEMC'09. IEEE 6th International. IEEE, pp. 2482–2485.

Holley, D. W., 2004. Understanding and using opc for maintenance and reliability applications. Computing & Control Engineering 15 (1), 28 – 31.
URL http://0-search.ebscohost.com.jabega.uma.es/login.aspx?direct=true&db=a9h&AN=12650871&lang=es&site=ehost-live&scope=site

Kalaitzakis, K., Koutroulis, E., Vlachos, V., 2003. Development of a data acquisition system for remote monitoring of renewable energy systems. Measurement 34 (2), 75 – 83.
URL http://www.sciencedirect.com/science/article/pii/S0263224103000253

Martínez Marchena I, Sidrach-de-Cardona M, M.-L. L., 2014. Framework for monitoring and assessing small and medium solar energy plants. Sol. Energy Eng.

Osterwald, C. R., 1986. Translation of device performance measurements to reference conditions. Solar cells 18 (3), 269–279.

Wang, L., Liu, K. H., Nov 2007. Implementation of a web-based real-time monitoring and control system for a hybrid wind-pv-battery renewable energy system. In: Intelligent Systems Applications to Power Systems, 2007. ISAP 2007. International Conference on. pp. 1–6.

Chapter 8

Conclusions and future work

The aim of this thesis was to present new models to forecast hourly solar radiation that could be helpful for electricity energy plants and eletricity distribution network administrators in order to know beforehand how much energy will solar power plants produce in a short-term.

Two models are proposed to forecast hourly solar radiation. The first one is based on using the cumulative probability distribution functions to identify different types of days. The K-means algorithm is proposed to cluster samples so that each cluster represents one type of day. A method is described to get the hourly solar radiation profile for a day taking as input only the daily clearness index (K_d) of that day. Results show a good performance as the energy error is 10.5% for all the data set, and a 5% for 57% of the energy. As a result of this analysis, it is possible to conclude that all estimated cumulative probability distribution functions can be summarized using six different types of curves. The results indicate that it is possible to use the obtained daily profiles of hourly solar radiation distribution to forecast the hourly values of this variable.

In the second proposed model, a new approach based on the use of different data mining techniques to model and to improve the next-day prediction of hourly global solar radiation has been developed. Modeling and prediction are conducted in two different phases and using different data mining techniques in each one. This second model introduces a new variable, which is related to the daily radiation profile and the daily clearness index. This variable is calculated for everyday and then K-means is applied to separate samples into clusters to identify different types of days.

A clustering algorithm is used to identify how many different day types there are in the first phase, and in the second phase, different classification algorithms are combined with regression algorithms to obtain the parameters to forecast hourly global solar radiation with four method combinations: decision trees, artificial neural networks, support vector machine regression and support vector machine classification.

The classification algorithms are used to estimate the cluster to which the day belongs to. Regression algorithms are used to estimate the daily clearness index. These two estimated parameters and the proposed centroids are used to predict the hourly global solar radiation. Four different types of daily profiles have been estimated. These profiles correspond to clear (cloudless), overcast (variable among the day), overcast mainly in the morning and overcast mainly in the afternoon.

Regarding the results obtained in Experiment 1, when the independent variables are the values of the meteorological parameters for the previous day, the $rMAE$ value for the best model is 16.7% and the $RMSE$ value 23.5%. The results obtained for $rMAE$ and $RMSE$ are 15.2% and 22.9% respectively for Experiment 2, when the independent variables are the forecasts of meteorological variables for the same day to be forecasted except for the value of daily clearness index that corresponds to the previous day. These values are inside the benchmark previously proposed by several authors.

The proposed model to forecast hourly solar global radiation using the forecasted values of meteorological parameters has been used to assess photovoltaic facilities. This model has been integrated into a monitoring framework for solar photovoltaic plants. With the solution proposed applications that integrate a single tool monitoring photovoltaic systems connected to network with inverters or data acquisition systems from different technologies, meter reading and production analysis procedures, evaluation, fault detection and alarm generation of the plant. This proposal enables evaluation of photovoltaic plants remotely using a single program, and without relying on software developed by inverters manufacturers.

Finally, a prototype application is presented in order to estimate hourly solar raditation values from real data taken from AEMET (Agencia Estatal de Meteorología), which shows the usefulness of the second proposed model. This application obtains data from Aemet station as well as Junta de Andalucía stations spread throughout the entire Region.

Future work includes testing models with larger data sets and from different locations, and testing with input sets that do not contain radiation information

(K_d) because there are locations that do not have solar radiation measurement equipment. Using more input variables can help to improve the forecasting accuracy because more variables may contain more information about atmosphere behaviour. Also, propose other machine learning models besides SVM, DT and ANN. Adaptative models are another type of model that could be investigated to improve results.

The developed or improved model could be incorporated to the existing monitoring system improving service to power plants administrators and users.

Appendices

Appendix A

Forecasting Web Tool

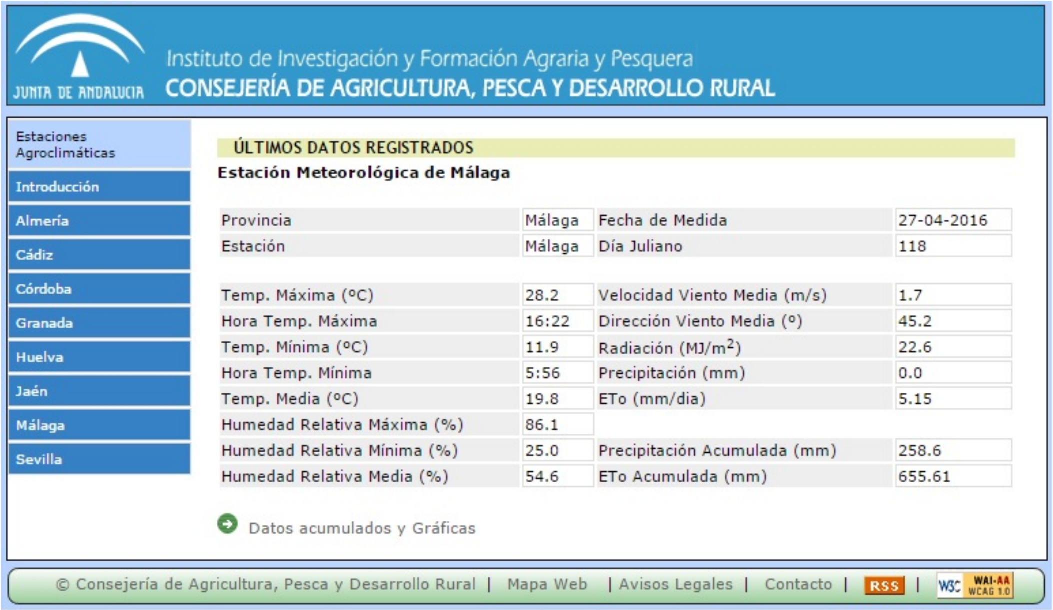
A.1 Forecasting Web Tool

A.1.1 Web application

A web application has been developed to obtain hourly forecasts using model presented in chapter 6. The application has been developed using *PHP*, *javascript* language and the LibSVM library which has a *PHP* extension of the SVM algorithms, (Chang and Lin, 2011). SVM-C and SVM-R are used because they achieved the best results against other methods combinations.

Forecast data (temperature and humidity for next day) from the spanish national weather service (*AEMET*) are used to perform the forecast process. These data are collected every day into a local database using a simple PHP script. The application collects information from several predefined meteorological stations in XML from the AEMET web site, which has an available API for this purpose.

Daily irradiance data are collected from Junta de Andalucia web site where a daily value of irradiance is available from several stations from all the region. From these data, K_d can be calculated using equation 4.17. The resulting values are stored in the same database mentioned above. An example of Junta de Andalucía's web site for the meteorological station of Málaga is shown in Figure A.1.


JUNTA DE ANDALUCIA
Instituto de Investigación y Formación Agraria y Pesquera
CONSEJERÍA DE AGRICULTURA, PESCA Y DESARROLLO RURAL
Estaciones Agroclimáticas
Introducción
Almería
Cádiz
Córdoba
Granada
Huelva
Jaén
Málaga
Sevilla
Datos acumulados y Gráficas
© Consejería de Agricultura, Pesca y Desarrollo Rural | Mapa Web | Avisos Legales | Contacto | RSS | W3C WAI-AA WCAG 1.0

ÚLTIMOS DATOS REGISTRADOS

Estación Meteorológica de Málaga

Provincia	Málaga	Fecha de Medida	27-04-2016
Estación	Málaga	Día Juliano	118
Temp. Máxima (ºC)	28.2	Velocidad Viento Media (m/s)	1.7
Hora Temp. Máxima	16:22	Dirección Viento Media (º)	45.2
Temp. Mínima (ºC)	11.9	Radiación (MJ/m^2)	22.6
Hora Temp. Mínima	5:56	Precipitación (mm)	0.0
Temp. Media (ºC)	19.8	ETo (mm/dia)	5.15
Humedad Relativa Máxima (%)	86.1		
Humedad Relativa Mínima (%)	25.0	Precipitación Acumulada (mm)	258.6
Humedad Relativa Media (%)	54.6	ETo Acumulada (mm)	655.61

Figure A.1: Junta de Andalucía Web Site for meteorological data

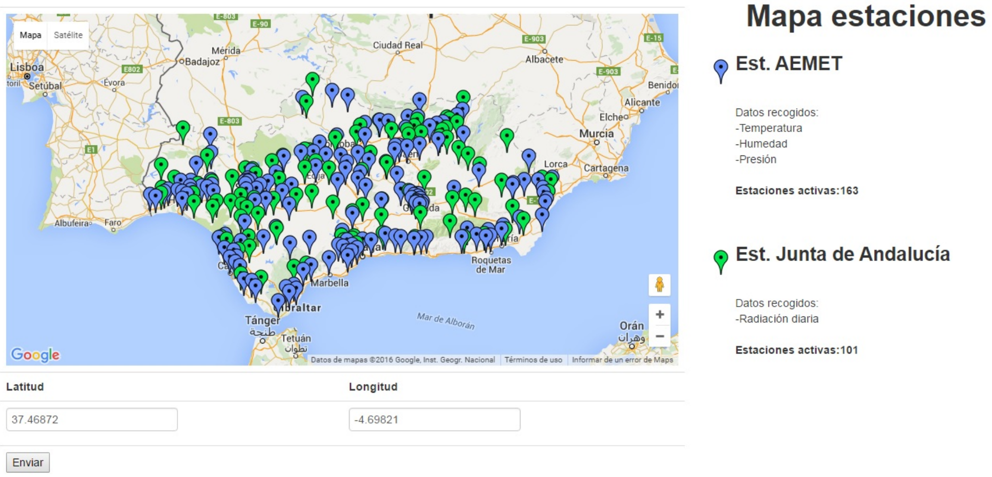
A model is trained using three years of data from specific location (Málaga), but the model can be used to forecast data for any location. The application works as follows:

- A map is shown to the user where he can select a desired location. The position of every meteorological station from AEMET and Junta de Andalucia are shown, AEMET stations are showed in blue and Junta de Andalucía stations are shown in green.

- Once the location is selected, the user clicks on the “Send” button. At this moment the application searches and selects the closest meteorological stations to the selected location using the geographical coordinates and euclidean distance.

- Next, the forecasted data of temperature, humidity and pressure from the closest station is selected and used as input in the previously trained model to forecast the cluster and K_d. Cluster is predicted using the SVM-C algorithm while K_d is predicted using the SVM-R algorithm.

- Once the two values are calculated, they are used to forecast the hourly solar radiation with the method described before: the centroid of he predicted cluster is selected and the predicted K_d is added to each of the eight values of the centroid, obtaining an estimation of K_h. Next, the hourly values of extraterrestrial radiation are multiplied by the corresponding estimated value of K_h and the results are the estimated values of hourly solar global radiation, expressed in $W/m2$.



Mapa estaciones

Est. AEMET

Datos recogidos:
-Temperatura
-Humedad
-Presión

Estaciones activas:163

Est. Junta de Andalucía

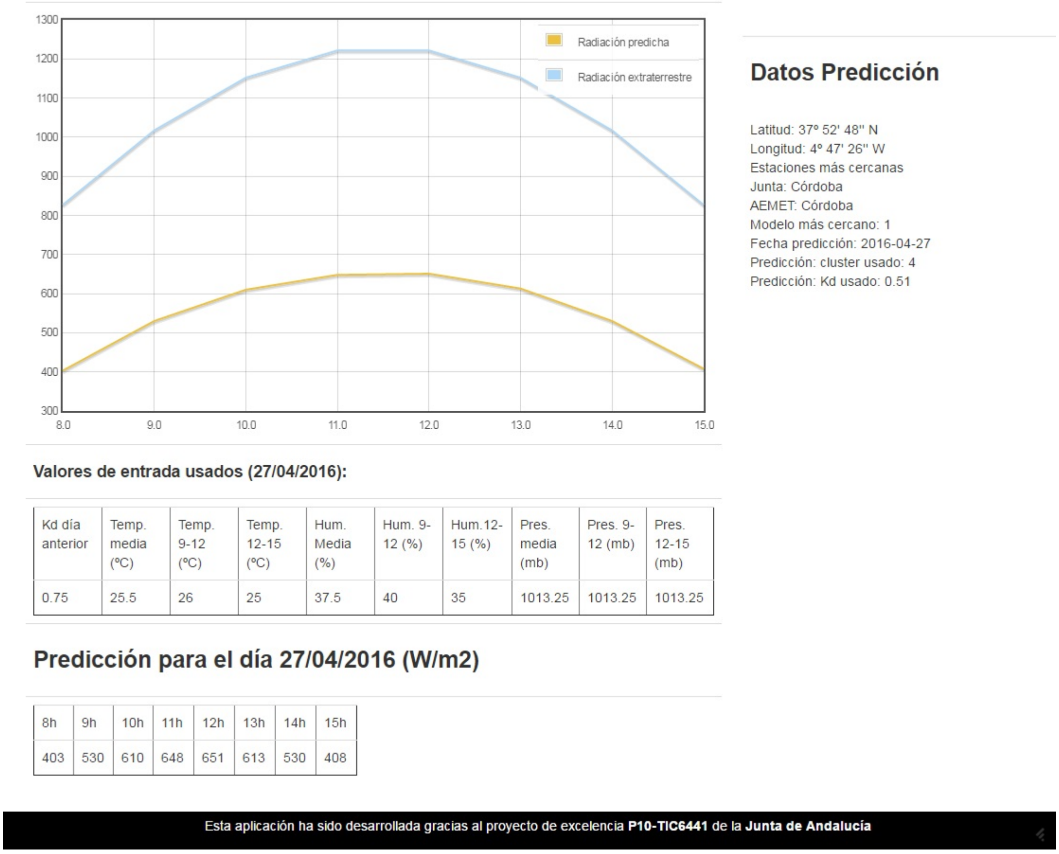
Datos recogidos:
-Radiación diaria

Estaciones activas:101

Latitud	Longitud
37.46872	-4.69821

Enviar

Figure A.2: Web application, step 1



Datos Predicción

Latitud: 37º 52' 48'' N
Longitud: 4º 47' 26'' W
Estaciones más cercanas
Junta: Córdoba
AEMET: Córdoba
Modelo más cercano: 1
Fecha predicción: 2016-04-27
Predicción: cluster usado: 4
Predicción: Kd usado: 0.51

Valores de entrada usados (27/04/2016):

Kd día anterior	Temp. media (ºC)	Temp. 9-12 (ºC)	Temp. 12-15 (ºC)	Hum. Media (%)	Hum. 9-12 (%)	Hum.12-15 (%)	Pres. media (mb)	Pres. 9-12 (mb)	Pres. 12-15 (mb)
0.75	25.5	26	25	37.5	40	35	1013.25	1013.25	1013.25

Predicción para el día 27/04/2016 (W/m2)

8h	9h	10h	11h	12h	13h	14h	15h
403	530	610	648	651	613	530	408

Esta aplicación ha sido desarrollada gracias al proyecto de excelencia **P10-TIC6441** de la **Junta de Andalucía**

Figure A.3: Web application, step 2

Bibliography

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27, software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Appendix B

AEMET data

B.1 Aemet example XML file

```
<root xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns:xsi="http://www.w3.org/2001/XMLSchema
    ↪ -instance" id="29067" version="1.0" xsi:noNamespaceSchemaLocation="http://www.aemet.es/
    ↪ xsd/localidades.xsd">
<origen>
<productor>
Agencia Estatal de Meteorología – AEMET. Gobierno de España
</productor>
<web>http://www.aemet.es</web>
<enlace>
http://www.aemet.es/es/eltiempo/prediccion/municipios/malaga-id29067
</enlace>
<language>es</language>
<copyright>
© AEMET. Autorizado el uso de la información y su reproducción citando a AEMET como autora de
    ↪ la misma.
</copyright>
<nota_legal>http://www.aemet.es/es/nota_legal</nota_legal>
</origen>
<elaborado>2016-01-07T20:24:02</elaborado>
<nombre>Málaga</nombre>
<provincia>Málaga</provincia>
<prediccion>
<dia fecha="2016-01-07">
<prob_precipitacion periodo="00-24"/>
<prob_precipitacion periodo="00-12"/>
<prob_precipitacion periodo="12-24">0</prob_precipitacion>
<prob_precipitacion periodo="00-06"/>
<prob_precipitacion periodo="06-12"/>
<prob_precipitacion periodo="12-18">0</prob_precipitacion>
<prob_precipitacion periodo="18-24">0</prob_precipitacion>
<cota_nieve_prov periodo="00-24"/>
<cota_nieve_prov periodo="00-12"/>
<cota_nieve_prov periodo="12-24"/>
<cota_nieve_prov periodo="00-06"/>
<cota_nieve_prov periodo="06-12"/>
<cota_nieve_prov periodo="12-18"/>
<cota_nieve_prov periodo="18-24"/>
<estado_cielo periodo="00-24" descripcion=""/>
<estado_cielo periodo="00-12" descripcion=""/>
<estado_cielo periodo="12-24" descripcion="Poco_nuboso">12</estado_cielo>
<estado_cielo periodo="00-06" descripcion=""/>
<estado_cielo periodo="06-12" descripcion=""/>
<estado_cielo periodo="12-18" descripcion="Poco_nuboso">12</estado_cielo>
<estado_cielo periodo="18-24" descripcion="Despejado">11n</estado_cielo>
<viento periodo="00-24">
```

```
<direccion/>
<velocidad/>
</viento>
<viento periodo="00-12">
<direccion/>
<velocidad/>
</viento>
<viento periodo="12-24">
<direccion>NO</direccion>
<velocidad>15</velocidad>
</viento>
<viento periodo="00-06">
<direccion/>
<velocidad/>
</viento>
<viento periodo="06-12">
<direccion>NO</direccion>
<velocidad>15</velocidad>
</viento>
<viento periodo="12-18">
<direccion>NO</direccion>
<velocidad>15</velocidad>
</viento>
<viento periodo="18-24">
<direccion>O</direccion>
<velocidad>15</velocidad>
</viento>
<racha_max periodo="00-24"/>
<racha_max periodo="00-12"/>
<racha_max periodo="12-24"/>
<racha_max periodo="00-06"/>
<racha_max periodo="06-12"/>
<racha_max periodo="12-18"/>
<racha_max periodo="18-24"/>
<temperatura>
<maxima>20</maxima>
<minima>12</minima>
<dato hora="06"/>
<dato hora="12">19</dato>
<dato hora="18">17</dato>
<dato hora="24">14</dato>
</temperatura>
<sens_termica>
<maxima>20</maxima>
<minima>12</minima>
<dato hora="06"/>
<dato hora="12">19</dato>
<dato hora="18">17</dato>
<dato hora="24">14</dato>
</sens_termica>
<humedad_relativa>
<maxima>75</maxima>
<minima>50</minima>
<dato hora="06"/>
<dato hora="12">55</dato>
<dato hora="18">65</dato>
<dato hora="24">75</dato>
</humedad_relativa>
<uv_max>2</uv_max>
</dia>
<dia fecha="2016-01-08">
<prob_precipitacion periodo="00-24">85</prob_precipitacion>
<prob_precipitacion periodo="00-12">0</prob_precipitacion>
<prob_precipitacion periodo="12-24">85</prob_precipitacion>
<prob_precipitacion periodo="00-06">0</prob_precipitacion>
<prob_precipitacion periodo="06-12">0</prob_precipitacion>
<prob_precipitacion periodo="12-18">0</prob_precipitacion>
<prob_precipitacion periodo="18-24">80</prob_precipitacion>
<cota_nieve_prov periodo="00-24"/>
<cota_nieve_prov periodo="00-12"/>
<cota_nieve_prov periodo="12-24"/>
<cota_nieve_prov periodo="00-06"/>
<cota_nieve_prov periodo="06-12"/>
<cota_nieve_prov periodo="12-18"/>
<cota_nieve_prov periodo="18-24"/>
<estado_cielo periodo="00-24" descripcion="Intervalos␣nubosos␣con␣lluvia␣escasa">43</
      ↪ estado_cielo>
<estado_cielo periodo="00-12" descripcion="Despejado">11</estado_cielo>
<estado_cielo periodo="12-24" descripcion="Intervalos␣nubosos␣con␣lluvia␣escasa">43</
      ↪ estado_cielo>
<estado_cielo periodo="00-06" descripcion="Despejado">11n</estado_cielo>
<estado_cielo periodo="06-12" descripcion="Despejado">11</estado_cielo>
<estado_cielo periodo="12-18" descripcion="Poco␣nuboso">12</estado_cielo>
```

```
<estado_cielo periodo="18-24" descripcion="Nuboso_con_lluvia_escasa">44n</estado_cielo>
<viento periodo="00-24">
<direccion>SE</direccion>
<velocidad>15</velocidad>
</viento>
<viento periodo="00-12">
<direccion>O</direccion>
<velocidad>15</velocidad>
</viento>
<viento periodo="12-24">
<direccion>SE</direccion>
<velocidad>15</velocidad>
</viento>
<viento periodo="00-06">
<direccion>C</direccion>
<velocidad>0</velocidad>
</viento>
<viento periodo="06-12">
<direccion>SE</direccion>
<velocidad>15</velocidad>
</viento>
<viento periodo="12-18">
<direccion>SE</direccion>
<velocidad>10</velocidad>
</viento>
<viento periodo="18-24">
<direccion>S</direccion>
<velocidad>10</velocidad>
</viento>
<racha_max periodo="00-24"/>
<racha_max periodo="00-12"/>
<racha_max periodo="12-24"/>
<racha_max periodo="00-06"/>
<racha_max periodo="06-12"/>
<racha_max periodo="12-18"/>
<racha_max periodo="18-24"/>
<temperatura>
<maxima>20</maxima>
<minima>11</minima>
<dato hora="06">11</dato>
<dato hora="12">19</dato>
<dato hora="18">17</dato>
<dato hora="24">16</dato>
</temperatura>
<sens_termica>
<maxima>20</maxima>
<minima>11</minima>
<dato hora="06">11</dato>
<dato hora="12">19</dato>
<dato hora="18">17</dato>
<dato hora="24">16</dato>
</sens_termica>
<humedad_relativa>
<maxima>90</maxima>
<minima>60</minima>
<dato hora="06">80</dato>
<dato hora="12">60</dato>
<dato hora="18">75</dato>
<dato hora="24">85</dato>
</humedad_relativa>
<uv_max>2</uv_max>
</dia>
<dia fecha="2016-01-09">
<prob_precipitacion periodo="00-24">100</prob_precipitacion>
<prob_precipitacion periodo="00-12">100</prob_precipitacion>
<prob_precipitacion periodo="12-24">0</prob_precipitacion>
<cota_nieve_prov periodo="00-24">1900</cota_nieve_prov>
<cota_nieve_prov periodo="00-12">1800</cota_nieve_prov>
<cota_nieve_prov periodo="12-24"/>
<estado_cielo periodo="00-24" descripcion="Intervalos_nubosos_con_lluvia">23</estado_cielo>
<estado_cielo periodo="00-12" descripcion="Nuboso_con_lluvia">24</estado_cielo>
<estado_cielo periodo="12-24" descripcion="Despejado">11</estado_cielo>
<viento periodo="00-24">
<direccion>O</direccion>
<velocidad>15</velocidad>
</viento>
<viento periodo="00-12">
<direccion>SO</direccion>
<velocidad>15</velocidad>
</viento>
<viento periodo="12-24">
<direccion>O</direccion>
<velocidad>15</velocidad>
```

```
</viento>
<racha_max periodo="00-24"/>
<racha_max periodo="00-12"/>
<racha_max periodo="12-24"/>
<temperatura>
<maxima>18</maxima>
<minima>8</minima>
</temperatura>
<sens_termica>
<maxima>18</maxima>
<minima>7</minima>
</sens_termica>
<humedad_relativa>
<maxima>95</maxima>
<minima>50</minima>
</humedad_relativa>
<uv_max>2</uv_max>
</dia>
<dia fecha="2016-01-10">
<prob_precipitacion periodo="00-24">70</prob_precipitacion>
<prob_precipitacion periodo="00-12">30</prob_precipitacion>
<prob_precipitacion periodo="12-24">65</prob_precipitacion>
<cota_nieve_prov periodo="00-24">1800</cota_nieve_prov>
<cota_nieve_prov periodo="00-12">1900</cota_nieve_prov>
<cota_nieve_prov periodo="12-24">1800</cota_nieve_prov>
<estado_cielo periodo="00-24" descripcion="Intervalos␣nubosos␣con␣lluvia␣escasa">43</
    ↪ estado_cielo>
<estado_cielo periodo="00-12" descripcion="Intervalos␣nubosos">13</estado_cielo>
<estado_cielo periodo="12-24" descripcion="Intervalos␣nubosos␣con␣lluvia␣escasa">43</
    ↪ estado_cielo>
<viento periodo="00-24">
<direccion>SO</direccion>
<velocidad>10</velocidad>
</viento>
<viento periodo="00-12">
<direccion>SO</direccion>
<velocidad>10</velocidad>
</viento>
<viento periodo="12-24">
<direccion>SO</direccion>
<velocidad>10</velocidad>
</viento>
<racha_max periodo="00-24"/>
<racha_max periodo="00-12"/>
<racha_max periodo="12-24"/>
<temperatura>
<maxima>16</maxima>
<minima>7</minima>
</temperatura>
<sens_termica>
<maxima>16</maxima>
<minima>7</minima>
</sens_termica>
<humedad_relativa>
<maxima>95</maxima>
<minima>55</minima>
</humedad_relativa>
<uv_max>2</uv_max>
</dia>
<dia fecha="2016-01-11">
<prob_precipitacion>95</prob_precipitacion>
<cota_nieve_prov>1900</cota_nieve_prov>
<estado_cielo descripcion="Muy␣nuboso␣con␣lluvia">25</estado_cielo>
<viento>
<direccion>SO</direccion>
<velocidad>15</velocidad>
</viento>
<racha_max/>
<temperatura>
<maxima>14</maxima>
<minima>10</minima>
</temperatura>
<sens_termica>
<maxima>14</maxima>
<minima>10</minima>
</sens_termica>
<humedad_relativa>
<maxima>95</maxima>
<minima>75</minima>
</humedad_relativa>
<uv_max>2</uv_max>
</dia>
<dia fecha="2016-01-12">
```

```
<prob_precipitacion>20</prob_precipitacion>
<cota_nieve_prov>1500</cota_nieve_prov>
<estado_cielo descripcion="Intervalos_nubosos">13</estado_cielo>
<viento>
<direccion>NO</direccion>
<velocidad>15</velocidad>
</viento>
<racha_max/>
<temperatura>
<maxima>14</maxima>
<minima>8</minima>
</temperatura>
<sens_termica>
<maxima>14</maxima>
<minima>6</minima>
</sens_termica>
<humedad_relativa>
<maxima>80</maxima>
<minima>55</minima>
</humedad_relativa>
</dia>
<dia fecha="2016-01-13">
<prob_precipitacion>15</prob_precipitacion>
<cota_nieve_prov/>
<estado_cielo descripcion="Intervalos_nubosos">13</estado_cielo>
<viento>
<direccion>SE</direccion>
<velocidad>10</velocidad>
</viento>
<racha_max/>
<temperatura>
<maxima>14</maxima>
<minima>6</minima>
</temperatura>
<sens_termica>
<maxima>14</maxima>
<minima>5</minima>
</sens_termica>
<humedad_relativa>
<maxima>85</maxima>
<minima>60</minima>
</humedad_relativa>
</dia>
</prediccion>
</root>
```

Appendix C

Publications

Journals (JCR)

- Modeling and forecasting hourly global solar radiation using clustering and classification techniques. Solar Energy. (accepted, pending publication).

Other refereed publications

- Jiménez Pérez, P., Martínez Marchena, I. Navarro Tapia, D. Piliougine Rocha, M., Mora López, L., Sidrach de Cardona, M. **Instalaciones fotovoltaicas conectadas a red: Sistemas de pequeña y mediana potencia. Monitorización conjunta y evaluación**. *Era solar: Energías renovables*, 173. pp. 30-35. SAPT Publicaciones Técnicas. S.L., 2013. ISSN 0212-4157.

Congress

- Jiménez, P. Mora-López, L. **Modeling daily profiles of solar global radiation using statistical and data mining techniques**. Thirteenth International Symposium on Intelligent Data Analysis, 30th October and 1st November, 2014, Leuven, Belgium. *Lecture Notes in Computer Science.* Volume 8819, 2014, pp 155-166.

- Jiménez, P.; Mora-López, L.; Sidrach-de-Cardona, M. **Solar global daily irradiation short-term forecasting using data mining techniques**. 3rd International Congress on Natural Sciences and Engineering, Tokyo, 7-9 Mai, 2014, Japan.

- Jiménez-Pérez, Pedro Francisco; Martínez-Marchena, Ildefonso; Navarro-tapia, Diego; Piliougine-Rocha, Michel; Mora-Lopez, Llanos; Sidrach De Cardona-Ortin, Mariano . **Experiencias en la monitorización de sistemas fotovoltaicos conectados a red**. XV Congreso Ibérico y X Congreso Iberoamericano de Energía Solar, VIGO, ESPAÑA, junio 2012.

Apéndice D

Resumen de la tesis

D.1 Introducción

A medida que la sociedad progresa, las necesidades de energía crecen, cada vez tenemos más aparatos electrónicos que consumen electricidad, viajamos más lejos y tenemos bienes más sofisticados y producirlos requiere más energía. Dispositivos que tradicionalmente funcionaban con otras fuentes de energía ahora funcionan con electricidad, por ejemplo vehículos y cocinas. La energía es un recurso estratégico y los países siempre quieren ser energéticamente independientes.

Integrar la generación de energía en las ciudades y los lugares donde se consume es interesante para poder minimizar el coste de infraestructuras y pérdidas en el transporte. Las energías renovables suelen ser fáciles de integrar en los entornos donde se consume energía porque provocan menos problemas de polución, ruido, espacio, etc.

Las fuentes de energía tradicionales tienen serios problemas tales como que son limitadas, producen contaminación y algún día se agotarán, de manera que la humanidad está obligada a encontrar fuentes de energía limpias y renovables. La sociedad es cada vez más consciente de su responsabilidad con el medio ambiente y es por eso que vivimos una época de grandes avances científicos. Las energías renovables son limpias, inagotables y se están desarrollando rápidamente gracias a iniciativas como el protocolo de Kyoto. Las energías renovables incluyen: eólica, hidroeléctrica, solar, biomasa, biocombustible, etc.

La cantidad de plantas de energía solar se ha incrementado notablemente en los últimos años debido a la necesidad de usar energías limpias que reduzcan las emisiones de efecto invernadero, políticas de apoyo a estas tecnologías, el aumento de la eficiecia y el abaratamiento de estos sistemas. El aumento del número de estos sistemas hace necesario el desarrollo de mejores herramientas para integralos en el sistema eléctrico tradicional.

Las renovables continuaron su avance en el año 2014 a pesar del desplome del precio del petróleo, con los mayores avances en energía hidroeléctrica, solar y eólica. Estas dos últimas habiendo sufrido abaratamientos significativos. En países en vías de desarrollo las renovables ofrecen una gran oportunidad de acelerar la transición hacia servicios energéticos modernos y garantizar el acceso a la energía, aunque los precios y la financiación suponen aún una gran barrera.

A pesar del uso creciente de energía, las emisiones de CO_2 permanecieron estables durante 2014 debido al aumento de fuentes renovables y a las mejoras en eficiencia. Se estima que las renovables proporcionaron el 19.1 % de toda la energía consumida en el mundo en 2013 y continuó aumentando en 2014. La caída de los costes ha hecho que la energía fotovoltaica pueda competir con los combustibles fósiles. En 2014 se instalaron 40 GW de energía fotovoltaica para alcanzar un total de 177 GW instalados globalmente. 2014 también fue testigo de un gran aumento de instalaciones fotovoltaicas de tipo residencial, donde los usuarios producen la energía que consumen y venden el excedente.

El tiempo de retorno de la inversión para sistemas fotovoltaicos depende de la localización geográfica: en el norte de Europa ronda los 2.5 años mientras que en el sur ronda 1.5 años, dependiendo de la tecnología instalada. La eficiencia de los inversores actualmente es muy alta, la cuota de mercado actual es del 50 % para inversores de pequeño y mediano tamaño, 48 % para inversores de gran tamaño y un 1.5 % para micro-inversores.

En España, la demanda de energía eléctrica aumentó en 2015 después de cuatro años de descenso, aunque la producción de renovable disminuyó debido a la caída de producción de energía hidroeléctrica. La capacidad instalada creció un 5 % para hidroeléctrica y un 0.5 % para fotovoltaica hasta alcanzar el 4.3 %. La producción de energía fotovoltaica aumentó un 0.8 % respecto al año 2014 hasta alcazar el 3.1 % del total.

En el régimen especial del mercado eléctrico español, los administradores de plantas productoras de energía deben comunicar al distribuidor una estimación de la producción para el día siguiente. Si esa estimación difiere de la producción real en más de un 5 %, se aplica una penalización. Además de la previsión de producción, las plantas de energía de más de 2 MWp deben proveer mecanismos para ayudar al sistema eléctrico contra huecos de tensión (BOE, 2010), estas medidas ya fueron antes aplicadas a las plantas eólicas (BOE, 2007).

Los principales retos a los que se enfrenta la energía solar fotovoltaica son la gestión de la demanda de energía, el desarrollo de sistemas de baterías y la previsión de la producción de energía. Las actividades humanas a lo largo del día requieren energía en diferentes momentos, los sistemas de autoconsumo aumentan el beneficio de los sistemas fotovoltaicos al tiempo que disminuyen el estrés del sistema de distribución. El autoconsumo puede ser mejorado con el uso de baterías. Además, en instalaciones de autoconsumo se suelen detectar cambios en el patrón de consumo de energía por parte de los usuarios (Luthander et al., 2015) La gestión de la demanda se puede mejorar de forma manual (usuarios conectan determinados aparatos a ciertas horas) o automática (con medidores, limitadores de carga y dispositivos inteligentes). El pricipal problema de las baterías actuales es su coste, su capacidad y su tendencia a auto-descargarse en periodos largos. Además, para evitar el estrés y un ciclo de vida corto la capacidad de la batería debe ser mayor de lo que realmente se usa. Un sistema de autoconsumo puede mejorar hasta un 24 % cuando se usan baterías como medio de almacenamiento.

Las plantas fotovoltaicas conectadas a la red eléctrica deben trabajar en colaboración con otros sistemas (nucleares, gas, carbón, hidroeléctricas, eólicas, etc.) por lo que saber cuanta energía son capaces de producir es de vital importancia para mantener la estabilidad del sistema y trabajar en un entorno competitivo. La producción de energía con tecnologías tradicionales (nuclear, gas, carbón) se puede controlar fácilmente pero en el caso de la fotovoltaica o la hidroeléctrica ello depende de factores meteorológicos, no controlables por el ser humano. Es por ello que el desarrollo de sistemas de predicción de la producción es fundamental para estas fuentes de energía. Los problemas más importantes a la hora de predecir la energía solar que pueden producir los sistemas fotovoltaicos son: la predicción de la radiación solar a corto plazo y el uso de modelos adaptativos que aprendan durante la operación del sistema.

D.2 Estado del arte

La predicción del rendimiento de sistemas que usan la radiación solar como recurso energético requiere, por una parte, el uso de valores predichos de las condiciones de funcionamiento reales (parámetros meteorológicos) y, por otra parte, usar modelos que permitan estimar el rendimiento de esos sistemas a partir de los parámetros predichos. Estimar la energía producida por plantas solares es difícil debido a la dependencia de variables meteorológicas, como la radiación solar y la temperatura. La radiación solar es un parámetro intermitente (sucesión día-noche), tiene una componente estacional (posición relativa Sol-Tierra) y un comportamiento estocástico. Los modelos de predicción deben ser capaces de recoger estas tendencias y reproducir la componente no determinista.

Los primeros intentos de predecir la radiación solar tuvieron lugar hace más de 20 años (Jensenius and Cotton, 1981) utilizando la técnica *Model Output Statistics (MOS)*, en los años siguientes sólo hubo pequeñas variaciones del método *MOS* en (Heck and Takle, 1987) y (Jensenius, 1989). Desde entonces numerosos trabajos han sido presentados para predecir la radiación solar, algunos basados en modelos físicos y otros asumiendo que la radiación solar se puede predecir como cualquier otro proceso. Los métodos estadísticos se pueden dividir en dos categorías: modelos estadísticos y modelos de minería de datos.

Los modelos estadísticos asumen que los datos tiene una estructura interna que puede ser identificada utilizando las funciones de autocorrelación y autocorrelación parcial. Los métodos ARMA y ARIMA han sido ampliamente usados, por ejemplo (Brinkworth, 1997), (Bartoli et al., 1983), (Aguiar et al., 1988), (Graham et al., 1988), (Aguiar and Collares-Pereira, 1992) y (Mora-López and de Cardona, 1998) proponen diferentes modelos para series horarias y diarias de índide de transparencia. (Mora-Lopez and Sidrach-de Cardona, 1998) propone un método ARMA(1,1) para generar series horarias de radiación global usando como único parámetro de entrada el valor medio mensual de la radiación diaria global.

Los modelos de minería de datos no necesitan asumir nada respecto a los datos. En (Sfetsos and Coonick, 2000) se presenta un nuevo método para predecir la radiación global horaria en superficie horizontal basado en la utilización de *Redes Neuronales Artificiales (ANN)* y un *Esquema Adaptativo de Inferencia Neuro-difusa*. Inicialmente sólo se usa una variable de entrada y luego se van añadiendo más para mejorar los resultados. La red neuronal consigue mejorar el error RMS un 74 % sobre el modelo de persistencia.

En (Perez et al., 2007) se presenta un modelo basado en la predicción de la cobertura del cielo y se compara con medidas tomadas en tierra y datos de satélite. Se consigue un $rMBE$ del 2 % y un $RMSE$ del 35 % para un horizonte de predicción de 4 horas. En (Mellit and Pavan, 2010) se propone un método práctico para predecir la radiación solar para las próximas 24 horas usando el valor medio diario de la radiación y la temperatura del aire. Se obtienen unos valores de MBE y $RMSE$ de 32 % y 67 % respectivamente. Un modelo de predicción de radiación solar a medio plazo es presentado en (Marquez and Coimbra, 2011) adoptando variables meteorológicas predichas por el *US National Weather Service (NWS)* como entradas de una red neuronal (ANN). Se utilizan varios métodos para seleccionar las variables de entrada más influyentes que resultan ser la cobertura de nubes, la probabilidad de lluvia y las temperaturas máxima y mínima. El error $rRMSE$ oscila entre el 15 % y el 22 % para diferentes modelos construidos. En (Wang et al., 2012) se propone un modelo de red neuronal usando *Statistical Feature Parameters (ANN-SFP)* para predicción de radiación solar a corto plazo, donde el vector de entrada es construido con varios parámetros estadísticos de irradiación y temperatura ambiente. La red neuronal tiene 24 salidas, una para cada hora del día.

Otro enfoque utilizado para predecir la radiación solar es el que propone utilizar imágenes que muestran la evolución de las nubes en el cielo, ya que las nubes son el principal factor de los cambios de radiación solar aparte de la posición relativa de la Tierra y el Sol. Con esta técnica normalmente se obtienen buenos resultados a corto plazo, entre 30 minutos y 6 horas. Las imágenes tomadas desde tierra suelen tener mucha más resolución espacial y temporal comparadas con las imágenes de satélite pero están limitadas a un emplazamiento concreto. Los métodos de este tipo se basan en determinar la estructura actual de las nubes a partir de las imágenes tomadas e intentar predecir la posición futura de dichas nubes. Las imágenes de satélite se utilizan en (Hammer et al., 1999) para predecir la radiación solar en la superficie terrestre. A partir de las imágenes se hace una extrapolación temporal del desarrollo de las nubes y se predice la radiación solar con entre 30 y 120 minutos de antelación. Las imágenes se toman del satélite METEOSAT y el método muestra una cierta mejora respecto al modelo de persistencia. En (Zarzalejo et al., 2009) se presenta un ajuste estadístico entre un parámetro de radiación solar normalizado y el índice de nubosidad. El experimento se realiza con datos de 28 estaciones radiométricas de España. El $RMSE$ baja del 21 % al 17 % al añadir nuevas variables de entrada comparado con el modelo Heliosat-2. En (Aguiar et al., 2015) se utilizan redes neuronales artificiales para predecir la radiación solar usando imágenes de satélite y mediciones en tierra, para un horizonte de predicción de entre 1 y 6 horas. Las estaciones están las Islas Canarias y los resultados muestran un valor $RMSE$ de 15.3 % y 24.3 % para el horizonte de una hora para cada estación y de 22.6 % y 36.2 % para el horizonte de 6 horas.

Un método que usa imágenes tomadas desde tierra se presenta en (Chow et al., 2011) y se utiliza para predecir el comportamiento de las nubes y la radiación solar. Se toman imágenes cada 30 segundos, se genera un mapa de nubes y se calculan las sombras sobre la superficie. Este método presenta algunas desventajas: no se puede calcular la altura de las nubes y, además, como apuntan los autores, las imágenes quedan incompletas por el brazo que sujeta la cámara. Las pruebas muestran un 70 % de acierto en la predicción. En (Cazorla et al., 2008) se desarrolló un sistema de imágenes del cielo casero para la estimación de la cobertura de nubes. El sistema captura imágenes multiespectrales cada 5 minutos y las analiza con un método basado en redes neuronales y un algoritmo genético que se utiliza para encontrar la combinación óptima de parámetros de entrada; con ello se reducen a tres paramétros de entrada de los dieciocho iniciales.

Los modelos numéricos de predicción meteorológica están basados en ecuaciones dinámicas y pueden predecir las condiciones de la atmósfera con varios días de antelación. Las ecuaciones y entradas de estos modelos son discretizadas en una rejilla tridimensional que se extiende verticalmente sobre la superficie terrestre. Debido a que estos modelos son computacionalmente muy costosos de ejecutar, sólo unos pocos están en funcionamiento (Traunmüller and Steinmaurer, 2010). Normalmente se ejecutan de dos a cuatro veces al día, las condiciones iniciales se derivan de satélites, radares, estaciones de tierra, etc. y son procesadas e interpoladas en una rejillas 3D. La resolución es relativamente gruesa, con espacios en la rejilla de entre 40 y 90 kilómetros, para reducir el coste computacional.

Los modelos de mesoescala o modelos de área limitada cubren un área gegráfica limitada, típicamente desde 5 kilómetros hasta varios cientos. Algunos ejemplos son el *Weather Research and Forecast (WRF)*, el *Mesoscale Atmospheric Simulation System (MASS)*, el *Skiron-CENER* y *HIRLAM-CIEMAT* (Perez et al., 2013).

HIRLAM (HIgh Resolution Limited Area Model) es un modelo numérico de predicción meteorológica de corto alcance que predice directamente las siguientes variables: temperatura ambiente, componentes de viento, humedad, nubes, presión y la altura geopotencial. Las condiciones iniciales para el modelo son derivadas de observaciones directas y extrapolación de estas variables. En (Lara-Fanego et al., 2012) se evalúa la fiabilidad de las predicciones para tres días de antelación para la radiación global horizontal (GHI) e irradiancia normal directa proporcionada por el modelo atmosférico WRF de mesoescala para Andalucía. Para un horizonte de prediccón de 24 horas el error MBE es de un 2 % para cielos despejados y de un 18 % para cielos nubosos. El MBE para irradiancia normal directa sube hasta el 10 % y 75 % respectivamente en las mismas condiciones.

En (Kostylev et al., 2011) diferentes modelos de predicción para diferentes escalas de tiempo son analizados mostrando que los basados en imágenes por satélite pueden ofrecer mejores resultados para hasta 6 horas de antelación. Los modelos numéricos ofrecen un buen rendimiento para 6 horas de antelación en adelante.

Finalmente, entre los trabajos revisados encontramos dos de especial interés por estar realizados en localizaciones mediterráneas, al igual que este trabajo. En (Perez et al., 2013) datos de tres ciudades del sur de España (Córdoba, Granada y Huelva) son analizados usando los siguientes métodos de predicción: European Centre for Medium-Range Weather Forecast (ECMWF), WRF-UJAEN, un modelo basado en Weather Research Forecast (WRF) operado por la Universidad de Jaén, el sistema regional de predicción del tiempo Skiron operado en el Centro Nacional de Energía Renovable (CENER) de España y el modelo HIRLAM (High Resolution Limited Area Model) operado por la Agencia Estatal de Meteorología (AEMET). Los resultados muestran que el modelo ECMWF es el mejor con un $RMSE$ compuesto del 22 % y un MAE compuesto del 13 %.

En (Pierro et al., 2015) se evalúan dos métodos de postprocesado basados en Model Output Statistics (MOS). El primero (MOSRH) es un módelo físico que mejora la predicción de la concentración de vapor de agua en la atmósfera. El segundo, (MOSNN) está basado en algoritmos de aprendizaje estocástico y usa una red neuronal artificial. El MOSNN se usa para refinar la salida del MOSRH (se conectan en serie) y el modelo resultante se llama MOS en cascada. El modelo se evalúa con datos de las ciudades de Roma y Lugano. Para Roma, MOSRH rinde un poco mejor ($RMSE$ del 29 %) que el MOSNN (31 %). Para Lugano, MOSNN rinde mucho mejor que el MOSRH ($RMSE$ de 38 % y 47 % respectivamente). Para el MAE encontramos diferencias similares, con 18 % para MOSRH y 19 % para MOSNN en el caso de Roma, y 32 % para MOSRH y 27 % para MOSNN en el caso de Lugano.

D.3 Materiales y métodos

En este apartado se decriben los distintos modelos propuestos para analizar y predecir la radiación solar. Los métodos basados en minería de datos han empezado a usarse para predecir la radiación solar en los últimos años. La propuesta que se hace en este trabajo es usar tanto métodos estadísticos como modelos de minería de datos para caracterizar y predecir la radiación solar global horaria. La radiación a lo largo del día puede ser analizada usando funciones de distribución de probabilidad acumulada, y se propone la utilización de técnicas de agrupamiento y el test de Kolmogorov-Smirnov para determinar los diferentes tipos de días.

En este trabajo se propone la utilización de diferentes métodos de minería de datos (agrupamiento, clasificación y regresión) para predecir la radiación solar. Los métodos de clasificación permiten clasificar muestras en clases predeterminadas basándose en observaciones pasadas. Se propone usar Arboles de Decisión (Decision Trees, DT) y Máquinas de Soporte Vectorial (Support Vector Machines for Classification, SVM-C) como métodos de clasificación.

Para evaluar el error cometido en las predicciones se usan varios métodos, en las predicciones de valores continuos se estima el error en función de la diferencia entre el valor predicho y el valor observado mientras que para clasificación se cuenta el número de muestras correctamente clasificadas.

La función de distribución de probabilidad acumulada (Cumulative Probability Distribution Function, CPDF) describe la probabilidad de que una variable X con una distribución de probabilidad dada tenga un valor menor o igual que x:

$$F_X(x) = \Pr(X \leq x) \tag{D.1}$$

El test de Kolmogorov-Smirnov para dos muestras está basado en la CPDF y se puede usar para comparar una muestra con una distribución de probabilidad de referencia o bien comparar dos muestras. Se cuantifica la distancia entre la función de distribución empírica de la muestra y la función de distribución acumulada de la distribución de referencia, o bien entre las funciones de distribución empíricas de dos muestras.

La regresión lineal es el método más simple y más usado de todas las técnicas estadísticas para modelar la dependencia de una variable sobre una o varias variables. La regresión es un método por el cual una relación funcional del mundo real se decribe con un modelo matemático que puede ser usado para explorar, describir y predecir dicha relación. Normalmente las variables de entrada se llaman variables independientes y la variable de salida se llama variable dependiente. La relación se puede expresar como una ecuación:

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \tag{D.2}$$

donde Y_t es la variable dependiente, $X_1, X_2, ..., X_p$ son las variables independientes y $\beta_0, \beta_1, \beta_2, ..., \beta_p$ son los coeficientes de influencia de las variables independientes sobre la dependiente. Determinar los valores de $\beta_0, \beta_1, \beta_2, ..., \beta_p$ es un problema de minimización.

Las técnicas de agrupamiento o *clustering* consisten en encontrar grupos de entidades con características similares dentro de cojuntos de datos, con los objetivos de estructurar los datos y poder sacar conclusiones sobre ellos. En este trabajo se usa la técnica de K-means, propuesta por MacQueen (MacQueen, 1967). Esta técnica consiste en dividir el conjunto de datos en K grupos o clusters y cada muestra pertenece al grupo con la media más cercana a la muestra. Es un método de aprendizaje no supervisado, sólo se proporciona el número de grupos deseados y el método hace el resto. El algoritmo se basa en examinar una o más características de las muestras y agruparlas según esas características. Para comparar las muestras se utiliza una medida que da la distancia entre un par de muestras, normalmente la distancia euclídea. Las medias de las muestras de cada grupo se llaman centroides. Al principio se eligen K centroides y se calculan los grupos (cada muestra va al grupo del centroide más cercano), luego se recalculan los centroides para cada grupo y así sucesivamente. El ciclo acaba cuando no se producen cambios en los grupos (ninguna muestra cambia de grupo). La elección de los centroides iniciales puede marcar el resultado final del algortimo por lo que existen varios métodos para elegir esos centroides iniciales.

Las redes neuronales artificiales están inspiradas en el sistema nervioso de los animales y es un tipo supervisado de paradigma del aprendizaje computacional. Una red neuronal artificial consiste en un número de neuronas interconectadas para producir una salida a partir de una entrada dada, aunque el diseño de la red depende del problema tratado. La red debe ser primero entrenada (aprendizaje supervisado) y luego puede ser usada para clasificar muestras o predecir valores. Normalmente las redes están organizadas en tres capas, la capa de entrada (una neurona por cada entrada), la capa oculta y la capa de salida (con una neurona por cada valor de salida). Una red neuronal se define normalmente por tres parámetros: el patrón de interconexión entre neuronas, el proceso de aprendizaje para los pesos de las conexiones y la función de activación de cada neurona (lineal, sigmoide, gausiana, etc). El método de aprendizaje más utilizado es el algoritmo de Levenberg-Marquardt.

Las Máquinas de Soporte Vectorial (Support Vector Machines, SVM) son un modelo de aprendizaje supervisado introducido por (Berthold and Hand, 2003) con la idea de ser un sistema capaz de separar las muestras con hiperplanos en un espacio de dimensión superior, capaz de analizar y reconocer patrones y que se pueda usar para análisis de clasificación y regresión. La implementación de una SVM está basada en dos pasos: (i) mapear las muestras de entrenamiento a un espacio de dimensión superior en el que las muestras son linealmente separables y (ii) determinar el hiperplano de separación óptima que maximiza el margen (la distancia de separación entre puntos de distintas categorías). En este trabajo se usa la implementación SVM de Chih-Chung Chang and Chih-Jen Lin (Chang and Lin, 2011), que soporta estos tres tipos de tareas: (i) SVC: clasificación, dos clases y multi clase, (ii) regresión y (iii) SVM de una única clase. En este trabajo se usan SVM tanto para clasificación como para regresión.

Los árboles de decisión (Decision Tree, DT) son otro tipo de modelo de aprendizaje supervisado. Los DT se pueden describir como grafos dirigidos acíclicos donde todos los nodos excepto el raíz tienen un solo nodo padre y cada uno puede tener cero, uno o más hijos. Los nodos sin hijos se llaman nodos hoja. Los DT pueden ser usados para clasificación (conjunto de valores finitos) y para regresión (valores continuos). Cada nodo interno (no hoja) representa un test sobre un atributo y cada salto hacia otro nodo representa un resultado para el test. Cada nodo hoja representa una etiqueta de clase (clasificación) o un valor de salida (regresión).

Las métricas usadas para medir la precisión de los modelos propuestos son varias:

- MAPE (Mean Absolute Percentage Error) es el error medio absoluto en porcentaje.
- MSE (Mean Square Error) es el error cuadrático medio.
- RMSE (Root Mean Square Error) es la raíz del error cuadrático medio.
- MAE (Mean Absolute Error) es el error medio absoluto.
- RMAE (Relative Mean Absolute Error) es el error medio absoluto relativo.

Para probar los modelos de predicción propuestos son necesarios datos meteorológicos consistentes y precisos. Estos datos se pueden obtener de servicios web, estaciones meteorológicas, etc. Los datos obtenidos de servicios web suelen estar libres de errores, en cambio, los que se obtienen directamente de estaciones meteorológicas están disponibles con más rapidez. En este trabajo se han utilizado datos de tres fuentes distintas: una estación meterológica propia, el servicio OpenWeatherMap y el servicio de la Agencia Estatal de Meteorología.

En el tejado de la ETSI de Informática de la Universidad de Málaga hay una estación meteorológica disponible, con instrumentos capaces de medir radiación solar global, temperatura, presión y humedad. Es un sistema modular, con un módulo principal que lee los datos de los módulos específicos para medir cada parámetro. El sistema proporciona ficheros de datos CSV que son fáciles de procesar. Los datos usados de esta estación corresponden al periodo entre octubre de 2010 y diciembre de 2013. Los datos son preprocesados para eliminar valores no deseados tales como fallos de medición, valores fuera de rango, etc.

OpenWeatherMap es un servicio que pretende hacer la información meteorológica disponible para todo el mundo de forma gratuita. Dispone de una API a través de la cual se puede acceder a varios tipos de datos diferentes (predicciones, observaciones, históricos). Hay disponibles predicciones con un horizonte de 5 días y con datos cada 3 horas. Los datos se pueden descargar en formato XML o en JSON.

La Agencia Española de Meteorología (AEMET) es el servicio oficial de meteorología de España. El sitio web de la agencia suministra datos tanto observados como de predicciones. Los datos de predicciones se pueden descargar en formato XML para su procesamiento e incluyen predicciones para un horizonte de 5 días de los siguientes parámetros: temperatura, humedad, probabilidad de lluvia e índice de cobertura de nubes. Las predicciones se dan con un inervalo de 6 horas. En este trabajo son de interés los datos de predicción para el día siguiente; estos datos se guardan en una base de datos desarrollada para tal efecto.

D.4 Fundamentos de la radiación solar

Los ángulos solares son útiles para saber cuanta radiación solar llega a la parte exterior de la atmósfera terrestre y, en conjunción con el índice de transparencia, sirven para estimar cuanta radiación llega a la superficie terrestre en un momento dado. Los ángulos solares incluyen el ángulo diario (depende del día del año), declinación (es el ángulo entre el eje de rotación de la Tierra y el plano perpendicular que cruza el centro de la Tierra), el ángulo horario (que es positivo durante la mañana, se reduce a cero a medio día y pasa a ser negativo por la tarde), el ángulo acimutal (es el ángulo entre la posición relativa del Sol y el sur geográfico), elevación (es el ángulo entre la línea que conecta el Sol con nuestra posición y su proyección en la superficie terrestre) y el ángulo cenital (es el complementario de la elevación).

Se define el concepto de radiación solar extraterrestre como la radiación recibida fuera de la atmósfera o la radiación en la superficie terrestre si no hubiera atmósfera. Se puede calcular usando la constante solar y la distancia entre la Tierra y el Sol. La constante solar se define como la cantidad de energía solar recibida por unidad de tiempo y de superficie para la distancia media Tierra-Sol. Se le da un valor de 1366.1 Wm^{-2} (Gueymard, 2004).

La series horarias son usadas como base para el estudio y análisis de las características y el comportamiento de la radiación solar y se usan como entrada para los algoritmos de aprendizaje automático para predecir la radiación global. Los datos recogidos por estaciones meteorológicas pueden ser grabados en diferentes peridos de tiempo, por ejemplo cada minuto, cada diez minutos, etc. Para obtener series horarias es necesario hacer antes un preprocesamiento. Es deseable que las series de datos no muestren ninguna tendencia para que el modelo de predicción pueda concentrarse en el comportamiento estocástico de la serie y evitar tener que predecir, por ejemplo, diferentes niveles de radiación en verano y en invierno.

Para eliminar las tendencias estacionales de las series se pueden utilizar métodos como las medias móviles usando funciones de Fourier o modelos físicos como el modelo de "cielo claro", capaz de estimar la irradiancia recibida en la superficie terrestre. El método usado en este trabajo para eliminar las tendencias en las series en el *índice de transparencia*, comúnmente usado para eliminar las tendencias diarias y estacionales de las series de datos de radiación solar. El índice de transparencia se calcula como el cociente entre la radiación global horizontal y la radiación extraterrestre para un lugar y momento dados:

$$K_t = \frac{G_t}{G_{t,0}}, \tag{D.3}$$

donde G_t es la radiacón solar global registrada para el momento t y $G_{t,0}$ es la radiación solar global extraterrestre para el mismo periodo.

Para cuantificar la precisión de los métodos de predicción de radiación solar se pueden usar varias métricas. Algunas reflejan mejor los costes de error de predicción y otras dan una mejor idea del rendimiento relativo bajo diferentes condiciones. Además, el conjunto de datos para hacer las pruebas debe estar libre de errores de observación y se debe dividir en datos de entrenamiento y datos de test.

Mean Bias Error (MBE) caracteriza el balance entre sobrepredicción y subpredicción. Es la medida de error parcial más utilizada y se define así:

$$MBE = \frac{1}{N}\sum_{t=1}^{N}(I(t) - \hat{I}(t)) \tag{D.4}$$

donde $I(t)$ es la irradiancia medida en el momento t, $\hat{I}(t)$ es la irradiancia predicha en el momento t y N es el número de muestras en el conjunto de datos. En el caso de una predicción perfecta ($\hat{I}(t) = I(t)$) esta métrica devuelve cero pero también en situaciones donde el error positivo y negativo se cancela y suman cero.

El coeficiente de determinación R^2 es capaz de medir cómo los valores predichos predicen las tendencias en los valores medidos. Es una comparación entre la varianza de los errores y la varianza de los datos, y se define como:

$$R^2 = 1 - \frac{\sigma^2(\hat{I} - I)}{\sigma^2(I)} \tag{D.5}$$

Para una predicción perfecta $R^2 = 1$.

El $RMSE$ está relacionado con la desviación estándar de los errores y se calcula a partir de la siguiente expresión:

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\hat{I}(t) - I(t))^2} \tag{D.6}$$

El modelo de persistencia se usa como referencia para ver la mejora de los modelos probados. En este modelo se asume que el siguiente valor a predecir es igual al anterior; por ejemplo, se asume que el valor de cambio entre el Euro y el Dólar mañana será igual al valor de hoy. Para radiación solar se puede desarrollar un modelo de persistencia asumiendo que la radiación solar mañana a una hora determinada será la misma que hoy a la misma hora.

El modelo *Forecast Skill* sobre predicción a 24 horas propuesto en (Coimbra and Kleissl, 2013) trata de establecer un índice de comparación entre el modelo de persistencia y el modelo que se está evaluando. El índice *forecast skill* de mejora s se define como:

$$s = 1 - \frac{RMSE_{model}}{RMSE_{persistence}}. \tag{D.7}$$

Por definición el *Forecast Skill* de persistencia es cero, y valores por encima de cero indican la mejora que tiene el modelo sobre el de persistencia. Si el valor es negativo entonces el rendimiento del modelo es peor que el de persistencia.

D.5 Predicción de perfiles horarios de radiación global solar utilizando funciones de distribución de probabilidad acumulada

Se propone un modelo para predecir la radiación solar global horaria que usa técnicas estadísticas y de minería de datos. En un primer paso el índice de transparencia se calcula para cada valor registrado, después, las CPDF se estiman para cada día y estas funciones son agrupadas usando el algoritmo de K-means, creando grupos de diferentes tipos de días según el perfil de radiación solar diaria. El número de clusters se establece en 6 usando el test de Kolmogorov-Smirnov para comprobar que son diferentes.

Seguidamente, se propone una elección del cluster usando el valor de K_d, esto es, se elige el cluster a partir del valor de K_d, que es conocido en este caso. Para cada cluster se calcula el valor medio del *índice de transparencia* y la desviación estándar y se comprueba que la desviación media es baja para los valores de las horas centrales del día, que son las que tienen más radiación.

Una vez se ha seleccionado el cluster, ya se tiene el correspondiente perfil K_h del cluster, y es posible calcular el valor de la radiación solar horaria simplemente multiplicando el valor de la radiación extraterrestre corespondiente (valores conocidos que sólo dependen de la fecha y hora) por cada valor del perfil K_h.

Para medir la precisión del método se calcula el error relativo en energía, que se define con la suma de todas las diferencias en valor absoluto de la energía predicha y la energía observada, dividido por la suma de la energía observada; esto da una medida relativa a la energía total observada. El error para todo el conjunto de datos es de 10.5 % mientras que el modelo de persistencia usado como referencia ofrece un error de 20.6 %. Además, el error es de un 5 % para el 57 % de la energía recibida (clusters 5 y 6).

D.6 Modelado y predicción a corto plazo de la radiación global solar

El modelo que se propone en este apartado es capaz de predecir también para el día siguiente el *índice de transparencia* diario. Se utiliza también en esta propuesta el algoritmo de K-means para agrupar los días según el tipo, pero se usan menos clusters ya que se usa una nueva variable definida para caracterizar el perfil de radiación solar de cada grupo de días. El modelo está basado en las siguientes hipótesis:

- Hay un número limitado de tipos de días según cómo están distribuidos los valores horarios de radiación.
- El tipo de día depende del valor del índice de transparencia diario.
- El índice de transparencia diario está relacionado con los parámetros meteorológicos del día anterior y del día actual.

De nuevo el *índice de transparencia* se usa para eliminar las tendencias diarias y estacionales de las series pero con una mejora, a cada perfil de *índice de transparencia* (compuesto por 8 valores) se le resta el valor del *índice de transparencia* diario, de manera que los días con similares perfiles pero a distinto nivel quedan mucho más igualados. Los valores de esta variable se calculan así:

$$k^*_{h,d} = (k_{h,d} - k_d) \qquad \texttt{para } h = 8 \texttt{ hasta } 15, \tag{D.8}$$

A continuación se agrupan los perfiles diarios de esta variable $k^*_{h,d}$ utilizando K-means, con $K = 4$ asumiendo que hay cuatro tipos distintos de días: días despejados, días cubiertos, días que van de despejados a cubiertos y días que van de cubiertos a despejados. La distancia euclídea es la usada para el algoritmo de K-means. Como resultado todas las muestras se clasifican en algún cluster y se obtiene un centroide por cada cluster que representa el valor medio de $k^*_{h,d}$.

Para estimar la radiación solar horaria de un día es necesario estimar el K_d y el cluster correspondiente, y como entrada se utilizan los datos meteorológicos (humedad, presión, temperatura) y el K_d del día anteior. Además, se proponen dos procedimientos distintos para llevar a cabo el proceso, en el primero, K_d y cluster se estiman de forma paralela, mientras que en el segundo primeramente se estima el cluster, y luego se hace la estimación del K_d de forma independiente para cada cluster usando las observaciones de cada cluster por separado, con lo que se espera que el resultado mejore.

Para estimar el cluster se utiliza un modelo de clasificación; en este caso se comparan dos, *SVM-C (Support Vector Machines for Classification)* y *DT (Decision Trees)*. Igualmente para estimar el valor de K_d se utilizan dos modelos, *SVM-R (Support Vector Machines for Regression)* y *ANN (Artificial Neural Network)*. Al usar dos métodos para calcular cada una de las dos variables se obtienen cuatro combinaciones de métodos para ser comparadas: DT+ANN, DT+SVM-R, SVM-C+ANN y SVM-C+SVM-R. Para hacer los experimentos se usa el 80 % de las muestras como conjunto de entrenamiento y el resto como conjunto de test. Los experimentos se llevan a cabo con MATLAB versión 2014b. Dos procesos son llevados a cabo para comparar resultados, un proceso de modelado donde se usan las variables meteorológicas del mismo día y el K_d del día anterior, y un proceso de predicción donde todas las variables usadas son del día anterior.

En los resultados se puede apreciar que la mejor combinación de métodos es SVM-C + SVM-R, consiguiendo en el caso del modelado (datos meteorológicos del mismo día) un rMAE del 15.2 % cuando los clusters se procesan por separado y un valor de 16.7 % en la caso de predicción (datos meteorológicos del día anterior) también procesando los clusters por separado. Como se esperaba, procesar los clusters por separado para el calculo del K_d ofrece resultados ligeramente mejores.

D.7 Predicción y evaluación del rendimiento de instalaciones fotovoltaicas

En la parte final del trabajo se presenta un sistema de evaluación de plantas solares fotovoltaicas. Se describe una arquitectura basada en tecnología OPC capaz de integrar fácilmente equipos de diferentes fabricantes y con diferentes características. Basta desarrollar un pequeño software llamado servidor OPC que obtiene los datos del equipo eléctrico/electrónico a través de su protocolo particular y suministra esos datos en un protocolo estándar de OPC de datos instantáneos o históricos.

Se propone un algoritmo de evaluación de la producción de sistemas fotovoltaicos en el que se usan parámetros basados en una estimación de la potencia que se espera producir (basado en mediciones o predicciones de la radiación solar recibida) y la potencia realmente producida, estas dos variables son usadas para estimar una variable que representa la productividad diaria (*daily yield*) y el rendimiento (*performance ratio*). Los valores de *daily yield* son comparados y si la diferencia excede unos valores límites, se notifica un problema en el funcionamiento de la planta. La estimación de la energía producida esperada usando predicción de radiación solar puede ser muy útil en lugares donde no se dispone de equipos de medición de la radiación solar.

D.8 Conclusiones y trabajo futuro

En este capítulo se resumen las principales conclusiones que se han alcanzado en el trabajo así como las propuestas de trabajo futuro. Entre estas conclusiones, las más importantes son:

- Es posible predecir los perfiles de radiación global horaria para un día utilizando las funciones de distribución acumuladas y el índice de transparencia diario. Los errores en energía de estas predicciones son del 10.5 %.

- Se han identificado un total de cuatro perfiles diarios de radiación global horaria, que están relacionados con los distintos tipos de cielo cubierto observados: días totalmente despejados, días totalmente cubiertos, días con nubosidad por la mañana y días con nubosidad por las tardes.

- Es posible predecir los valores horarios de radiación global para el día siguiente utilizando sólo algunos parámetros meteorológicos del día actual. Estos parámetros están disponibles normalmente en las páginas web de los servicios de meteorología estatales. El error en la predicción es del 16 %.

- Es posible hacer estas predicciones utilizando las predicciones de parámetros meteorológicos que suministran los servicios meteorológicos y modelos de minería de datos.

- Los modelos desarrollados pueden ser utilizados para la evaluación de sistemas fotovoltaicos en los emplazamientos en los que no se dispone de estaciones meteorológicas. Para ello, se propone la utilización de los valores obtenidos en estaciones meteorológicas de las agencias estatales.

Como trabajo futuro se plantea la posibilidad de usar como parámetros de entrada a los modelos tanto los valores registrados en un día como las predicciones para el día siguiente de los parámetros meteorológicos. Esto podría mejorar las predicciones. También, utilizar datos de otros emplazamientos, de cara a generalizar los resultados obtenidos.

D.9 Apéndices

Finalmente, en los apéndices, se presenta una herramienta que toma los datos de entrada de la Agencia Estatal de Meteorología y de la Web de la Junta de Andalucía y es capaz de predecir el perfil de radiación global horaria para el día completo.

Bibliografía

Aguiar, L. M., Pereira, B., David, M., Díaz, F., Lauret, P., 2015. Use of satellite data to improve solar radiation forecasting with bayesian artificial neural networks. Solar Energy 122, 1309 – 1324.
URL http://www.sciencedirect.com/science/article/pii/S0038092X15005927

Aguiar, R., Collares-Pereira, M., 1992. T.a.g: A time dependent autoregressive gaussian model for generating synthetic hourly radiation. Solar Energy 49(3), 167–174.

Aguiar, R., Collares-Pereira, M., Conde, J., 1988. Simple procedure for generating sequences of daily radiation values using a library of markov transition matrices. Solar Energy 4. (3), 269–279.

Bartoli, B., Coluaai, B., Cuomo, V., Francesca, M., Serio, C., 1983. Autocorrelation of daily global solar radiation. Il nuovo cimento 40, 113–122.

Berthold, M. R., Hand, D. J. (Eds.), 2003. Intelligent Data Analysis: An Introduction, 2nd Edition. Springer Verlag.

BOE, 2007. Boletín oficial de estado. real decreto 661/2007, de 25 de mayo, por el que se regula la producción de energía eléctrica en régimen especial.

BOE, 2010. Bolet'ín oficial del estado. real decreto 1565/2010, de 19 de noviembre, por el que se regulan y modifican determinados aspectos relativos a la actividad de producción de energía eléctrica.

Brinkworth, B., 1997. Autocorrelation and stochastic modelling of insolation sequences. Solar Energy 19, 343–347.

Cazorla, A., Olmo, F. J., Alados-Arboledas, L., Jan 2008. Development of a sky imager for cloud cover assessment. J. Opt. Soc. Am. A 25 (1), 29–39.
URL http://josaa.osa.org/abstract.cfm?URI=josaa-25-1-29

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27, software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chow, C. W., Urquhart, B., Lave, M., Dominguez, A., Kleissl, J., Shields, J., Washom, B., 2011. Intra-hour forecasting with a total sky imager at the {UC} san diego solar energy testbed. Solar Energy 85 (11), 2881 – 2893.
URL `http://www.sciencedirect.com/science/article/pii/S0038092X11002982`

Coimbra, C., Kleissl, J., 2013. Solar Resource Assessment and Forecasting. Elsevier, Waltham, Massachusetts, Ch. Chapter 8: Overview of Solar-Forecasting Methods and Metric for Accuracy Evaluation, pp. 171–194.

Graham, V., Hollands, K., Unny, T., 1988. A time series model for kt with application to global synthetic weather generation. Solar Energy 40, 83–92.

Gueymard, C. A., 2004. The sun's total and spectral irradiance for solar energy applications and solar radiation models. Solar Energy 76 (4), 423 – 453.
URL `http://www.sciencedirect.com/science/article/pii/S0038092X03003967`

Hammer, A., Heinemann, D., Lorenz, E., Lückehe, B., 1999. Short-term forecasting of solar radiation: a statistical approach using satellite data. Solar Energy 67 (1–3), 139 – 150.
URL `http://www.sciencedirect.com/science/article/pii/S0038092X00000384`

Heck, P., Takle, E., 1987. Objective forecasts of solar radiation and temperature. Iowa State Journal of Research 62, 29–42.

Jensenius, J., 1989. Insolation forecasting. Solar Resources, MIT Press, Cambridge, 335–349.

Jensenius, J., Cotton, G., 1981. The development and testing of automated solar energy forecasts based on the model output statistics (mos) technique. In: 1st Workshop on terrestrial solar resource forecasting and on use of satellites for terrestrial solar resource assessment, Washington, DC.

Kostylev, V., Pavlovski, A., et al., 2011. Solar power forecasting performance–towards industry standards. In: 1st International Workshop on the Integration of Solar Power into Power Systems Aarhus, Denmark.

Lara-Fanego, V., Ruiz-Arias, J., Pozo-Vázquez, D., Santos-Alamillos, F., Tovar-Pescador, J., 2012. Evaluation of the {WRF} model solar irradiance forecasts in andalusia (southern spain). Solar Energy 86 (8), 2200 – 2217, progress in Solar Energy 3.
URL `http://www.sciencedirect.com/science/article/pii/S0038092X11000582`

Luthander, R., Widén, J., Nilsson, D., Palm, J., 2015. Photovoltaic self-consumption in buildings: A review. Applied Energy 142, 80–94.

MacQueen, J. B., 1967. Some methods for classification and analysis of multivariate observations. In: Fifth symposium on math, statistics, and probability. Berkeley, CA: University of California Press. p. 281–297.

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.
URL `http://www.sciencedirect.com/science/article/pii/S0038092X11000193`

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.
URL `http://www.sciencedirect.com/science/article/pii/S0038092X10000782`

Mora-López, L., de Cardona, M. S., 1998. Multiplicative arma models to generate hourly series of global irradiation. Solar Energy 63, 283–291.

Mora-Lopez, L., Sidrach-de Cardona, M., 1998. Multiplicative arma models to generate hourly series of global irradiation. Solar Energy 63 (5), 283–291.

Perez, R., Lorenz, E., Pelland, S., Beauharnois, M., Knowe, G. V., Jr., K. H., Heinemann, D., Remund, J., Müller, S. C., Traunmüller, W., Steinmauer, G., Pozo, D., Ruiz-Arias, J. A., Lara-Fanego, V., Ramirez-Santigosa, L., Gaston-Romero, M., Pomares, L. M., 2013. Comparison of numerical weather prediction solar irradiance forecasts in the us, canada and europe. Solar Energy 94, 305 – 326.
URL `http://www.sciencedirect.com/science/article/pii/S0038092X13001886`

Perez, R., Moore, K., Stackhouse, P., 2007. Forecasting solar radiation preliminary evaluation of an approach based upon the national forecast database. Solar Energy 81(6), 809–812.

Pierro, M., Bucci, F., Cornaro, C., Maggioni, E., Perotto, A., Pravettoni, M., Spada, F., 2015. Model output statistics cascade to improve day ahead solar irradiance forecast. Solar Energy 117, 99 – 113.
URL `http://www.sciencedirect.com/science/article/pii/S0038092X15002212`

Sfetsos, A., Coonick, A., 2000. Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. Solar Energy 68 (2), 169 – 178.
URL `http://www.sciencedirect.com/science/article/pii/S0038092X9900064X`

Traunmüller, W., Steinmaurer, G., 2010. Solar irradiance forecasting, benchmarking of different techniques and applications of energy meteorology. In: Proceedings of the EuroSun 2010 conference.

Wang, F., Mi, Z., Su, S., Zhao, H., 2012. Short-term solar irradiance forecasting model based on artificial neural network using statistical feature parameters. Energies 5 (5), 1355.
URL `http://www.mdpi.com/1996-1073/5/5/1355`

Zarzalejo, L. F., Polo, J., Martín, L., Ramírez, L., Espinar, B., 2009. A new statistical approach for deriving global solar radiation from satellite images. Solar Energy 83 (4), 480 – 484.
URL `http://www.sciencedirect.com/science/article/pii/S0038092X08002223`