POLITECNICO DI MILANO

Dipartimento di Elettronica, Informazione e Bioingegneria

MASTER DEGREE

in Automation and Control Engineering

Ensemble methods on improving quality of solar radiation prediction and their comparison with a commercial forecasting service

Candidato:

Dario La Carrubba

Relatore:

Prof. Luca Ferrarini

a.a. 2018/2019

Acknowledgements

«*Ogni corpo immerso parzialmente o completamente in un fluido riceve una spinta verticale dal basso verso l'alto, uguale per intensità al peso del volume del fluido spostato*». [Principio di Archimede, matematico e fisico siracusano, vissuto nel III secolo a.C.]

Nei miei ringraziamenti per questo fantastico percorso giunto ormai al termine, ho voluto citare uno dei principi più importanti mai scoperti da uno dei siciliani più grandi della storia. Nella mia vita io sono stato in più occasioni come quel *corpo immerso parzialmente o completamente in un fluido* di cui Archimede parla, il *corpo* senza il *fluido* cadrebbe in basso per effetto della gravità ma *riceve una spinta verticale dal basso verso l'alto, uguale per intensità al peso del volume del fluido spostato.* Volevo quindi ringraziare profondamente il mio "*fluido*" che mi ha permesso di raggiungere traguardi che forse neanche immaginavo. Gran parte di esso è indubbiamente la mia famiglia, per me sacra, formata da gente incredibile e ostinata più di me a farmi andare sempre più in alto.
Volevo ringraziare quindi mio Padre Antonio, esempio di vita, che con la sua dote di risolvere qualsiasi problema mi ha insegnato a fare altrettanto. Mia Madre Lucia, altro esempio di vita, che con il suo supporto morale e la sua tenacia mi ha permesso di prendere decisioni importanti e portare a termine qualsiasi obiettivo. Mio fratello Tomas, persona con il quale ho condiviso fortunatamente tutta la mia vita, persona che ha dato e fatto sempre il massimo per me. Queste persone hanno investito tutto ciò che hanno per la mia crescita a livello culturale, personale, professionale, e sono sicuro che lo faranno per sempre. La famiglia è una di quelle cose che non si possono comprare, nè scegliere, e io non potevo essere più fortunato.
Grazie di essere stati, essere ancora, ed essere in futuro il mio *fluido.*

Un ringraziamento va anche a tutta la gente che ho avuto l'onore di incontrare e con i quali ho condiviso mille esperienze, tra questi i CdB, la Cumacca, i PoliCeschi, tutti gli amici più stretti conosciuti in Erasmus, e anche quelli che ho frequentato a Milano.

Un ringraziamento al dott. Le Anh Dao, per avermi supportato in questo progetto di tesi, e per aver fatto accrescere le mie competenze e uno al Prof. Luca Ferrarini per l'opportunità che mi ha offerto e le esperienze professionali che ne sono derivate e che ne deriveranno.

Adesso un viaggio nuovo comincia, con un bagaglio nuovo e sempre lo stesso *fluido.*

"A vulisti a bicicletta? Ora 'ncuminza a pitaliari" [Un saggio siciliano]

Abstract

The increasing use of renewable energy sources necessitates accurate forecasting models for generation scheduling. Amongst the renewable sources, solar and wind have gained acceptance and are being increasingly used in distributed generation. The main problem with these sources is the dependence of their power output on natural environmental parameters which are difficult to predict. Along with the discussion, focusing on the solar power, this thesis addresses the problem of estimation of the Solar Radiation (SR). Different setups of the time series model, and with their combinations with the weather forecast services using ensemble methods have been evaluated in medium-term prediction of the day-ahead regional SR. Moreover, several considerations including Support Vector Machine methods are also adopted at this stage, to cluster data. At the end, the validation of the approach is performed by using a SR data from a meteorological station and its nearby meteorological service which both situated in Northern Italy.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The focus on renewable energy sources is increasing tremendously, which motivates studies concerning the integration of renewable sources like wind and solar into existing energy systems. When the conventional sources are used for power generation, the generation is predominantly controlled by the machine ratings and generation capacity of the plant. Hence, there is a lesser need for short term generation forecasting. But in case of power generation from renewable energy sources, the generation is uncertain because the weather is erratic, and the generation depends heavily on weather conditions. Thus, there is a dire need to build forecast models for the generation in order to have better generation scheduling. Therefore with greater penetration of renewable sources in power generation, the focus is shifting towards generation forecasting (1) (2) (3). Considering the advantages of solar energy as a sustainable energy, power prediction for photovoltaic installations is a decisive factor. Thus, development and research on solar power has been rising year by year. The predictions are used to optimize usage of the solar energy and provide reasonably accurate knowledge of the solar resource availability at any location (4). Since the generation of power from solar energy is very erratic due to its heavy dependence on weather, seasonal changes, geographical location, time of the day, orientation and position of panel, etc.,

the forecasting methods may not give uniformly efficient results for all regions. A ubiquitously efficient forecast system, minimizing errors on the behavioral patterns of wind and solar energy has become a major subject matter of study for researchers across the globe (5). Hence there is a need to critically examine the seasonality and the nature of the data to determine the models that can be satisfactorily used for prediction of solar power generation.

There are two ways to address the issue of solar prediction. One is by creating very complex models which best describe nature. Such models are used by Meteorological departments for weather forecasts. Computations are highly complex, and a very powerful computer is used to solve the differential equations involved. The other method is to use statistical data and predict solar irradiance to a lower accuracy as compared to the former method but with less computational requirements.

In this thesis has been studied a few methods to improve the performance of an existing very complex predictor using time series data of a measured solar irradiance to shrink the gap between the accuracy of the empirical methods (very complex and expensive) and statistical ones (less complex and cheaper).

1.2 Background Studies

In the past, various studies in field of solar energy were conducted. In an attempt to assess solar irradiance, multiple classes of models such as regressors in logs, seasonal autoregressive integrated moving average (SARIMA), transfer functions, artificial neural networks (ANN), and hybrid models have been tested and compared each other.

Some of these studies demonstrate that the success of SARIMA models is attributable to its ability to capture the cycles more effectively than the other methods (6). For example in (6) is shown the evidence of the efficiency of SARIMA models with respect to the others in evaluating the 60-minutely and 10-minutely averaged global horizontal irradiance related to 13 and 15 days periods in two winters and two summers. In (7) an attempt to estimate the global horizontal solar irradiance was done and the result of a comparison between a few models brought the evidence that neural networks or hybrid models in some cases can improve at very high resolutions on the order of 5 min while SARIMA models have strong ability to capture the diurnal cycle more efficiently.

It is intuitive that a strong factor that makes solar radiation change in time is the presence of the clouds above the area in which it is measured or predicted. Thus, the idea to involve in some way the contribution of the cloud in the prediction should be profitable. In (8) and (9) some experiment in such sense has been done and it is clearly expressed the benefits of using cloud coverage predictions for predicting the solar irradiance. In (9) a solution using a sky-image-based cloud estimation and a solar irradiance drop using back projection is studied and the accuracy of the prediction improved. In (10) many solar irradiance forecasting models have been developed, such as time series models like ARIMA, satellite data based models, sky images based model, ANN models and wavelet analysis based models. Here it is stated that for forecasting horizon from 5 min to 4 hours ARIMA presents the best accuracy and that cloud imagery and hybrid model can improve the forecasting when solar irradiance present strong variability.

For a long-term forecasting the authors in (11) developed new methods for forecasting appropriately the monthly climate data in Indonesia. In particular, they studied three main forecasting methods which are Autoregressive Integrated Moving Average (ARIMA), Feed Forward Neural Networks (FFNN), and an averaging method for ensemble ARIMA, FFNN, and ARIMA-FFNN. The results showed that ARIMA yields more accurate forecast in training datasets than other methods, whereas the best method in testing datasets is FFNN, and as a conclusion very complex methods do not necessarily produce more accurate forecast than simpler one. Thus, once again it is stated that ARIMA works better for short-term forecasting while a reliable method for medium/long-term forecasting could be a Machine Learning method. But, according to (12) and (13) (14) (15) (16) (17) (18) Machine Learning methods require often large amount of data, have high computational cost, and high dependence on feature selection.

Nowadays, researchers reserved more attention on trying to combine a few models to possibly exploit the benefits of each one. For example, in (12) a new ensemble method for day-ahead regional photovoltaic (PV) power forecasting with hourly resolution has been proposed. Predicting efficiently the photovoltaic power require a good prediction of solar radiation, which is one of the most influent variables that regards photovoltaic power. The method utilizes open weather forecast and power measurement data, this prediction method is processed within a set of historical data with similar meteorological data (temperature and irradiance), and astronomical date (solar time and earth declination angle). Further, clustering and blending strategies are applied to improve its accuracy in regional PV forecasting. The method introduces robustness comparing it with other models which are the

North American Mesoscale Forecast System, the Global Forecast System, and the Short-Range Ensemble Forecast. Thus, combining historical data-driven predictors with machine learning ones could bring benefits.

1.3 Aims and Objectives

Considering the background studies and the fact that the goal here is to improve the accuracy of a medium-term prediction provided by a forecasting service, the aim of this study has been set mainly to analyze and possibly discover the potentiality of a few version of SARIMA models, some hybrid models that tries to exploit even the available prediction of cloud coverage, and ensemble models in predicting the available solar irradiance with respect to the prediction provided by the same complex model of a forecasting service.

1.4 Thesis Structure

The thesis is organized as follows:

- Chapter 2 gives an outline of the nature of the data available.
- Chapter 3 gives a detailed explanation of the various time series models tested and their main properties are discussed.
- Chapter 4 describes the ensemble methods tested.
- Chapter 5 describes the Auto Regressive Integrated Moving Average with eXogenous input (ARIMAX) model with the introduction of a clustering process performed by a Support Vector Machine (SVM).

- Chapter 6 describes the method to evaluate the goodness of each model and shows the results comparing them with the one provided by the prediction of the forecasting service and each other.
- Chapter 7 gives some conclusion of the work and some possible future ones.

Chapter 2

Data description

The most important part of developing a good predictor is to understand the nature of the data, their properties such as quality, correlations with the variable to predict and statistical properties. Thus, choosing the model structure most able to exploit the available data.

The available data in this thesis are all regarding a northern zone of Italy which is Vipiteno, Trentino Alto Adige.

The available data are:

- Measurement of solar radiation [W/m^2]
- Prediction of solar radiation by a forecasting service [W/m^2]
- Prediction of cloud coverage % by a forecasting service

2.1 Solar radiation measured

Fundamental concepts

In the field of solar power supply the most significant measures are the energy delivered and the intensity. The *solar irradiance* is the *rate* at which the solar energy reaches a unit area at the earth's surface. The unit of measurement for solar irradiance is Watt per meter square [W/m^2]. The *solar irradiation* is

instead the *integration* or *summation* of the *solar irradiance* over a time period as follows:

$$H = \int_{t_0}^{t_1} I(t)dt$$

where $I(t)$ is the solar irradiance at time t, t_1 and t_0 are the two limits of a time interval in which the solar irradiation is computed. The most common units of measurement of solar irradiation H are Joule per meter square (J/m^2) or Watt-hours per meter square (Wh/m^2). (19)

Data available

The data are collected by a meteorological station situated in Vipiteno which is at 948 meters over the level of the sea and at the geographic coordinates 46°53′N 11°26′E.

The available measurements of solar radiation are for a period that goes from 1st January 2014 00:15 to 2nd August 2018 1:00 with a sample time of 15 minutes. (Figure 1)

The quality of the data is good considering that there are just a few brief missing periods in the order of some half-hours and a few values considered outliers considering the nature of the data itself.

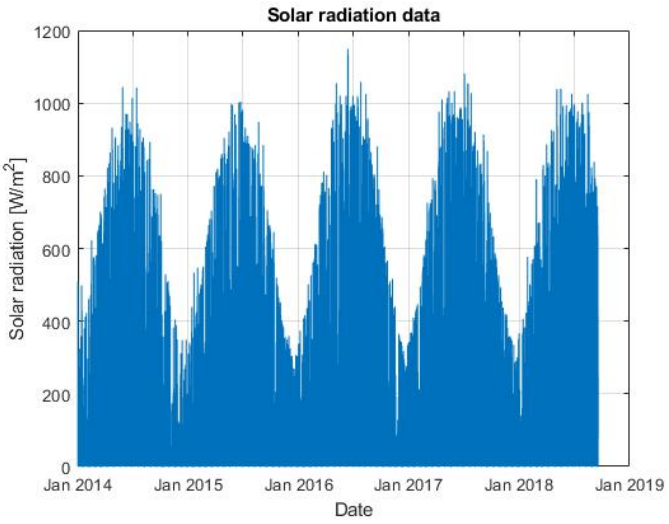

Figure 1 Solar radiaiton data

2.1.1 Missing data and outliers

Outliers

In this work, a value in the dataset has been considered *outlier* if:

- is different from 0 W/m² in the hours after the actual hour of sunset and before the actual hour of sunrise with a tolerance for that region and for the corresponding season
- is strongly higher than the typical maximum solar radiation in W/m² in that region and in that period of the year

To recognize every *outlier* in the data set an *ad-hoc function* was developed in MATLAB environment. A brief explanation of the mentioned function follows.

The function is called *sunsetsunrise_italy(date,value)* that takes as input the *date* and the *value* of a sample in the dataset, computes, for that date in the format *"dd/mm/yyyy/ hh:mm"* and based on the values of *latitude*, *longitude*

and *altitude* of the geographic location 46°53′N 11°26′E, the exact sunset and sunrise time, and a *Nan value* is substituted to the respective value if the following condition is fulfilled:

$$if\ [Date(t) \geq\ sunset.Date(t) +\ \delta] \wedge [Date(t) \leq sunrise.Date(t) - \delta] \wedge [value(t) > 0]$$

where δ is a tolerance value that in this case has been set to 1-hour. The function *sunsetsunrise_italy(date,value)* to compute the right sunset and sunrise time uses another function *sunrise(lat,lon,alt,tz,dte)* on the MATLAB library. A detailed explanation can be found reading (20) and (21).

Missing values

The dataset, as mentioned, presents some missing values caused by various phenomena such as equipment malfunction and interruptive maintenance of the sensors.

Thus, before studying the main statistical properties of the data, the entity of missing samples has been evaluated.

The missing data are in most of the cases of a few values in a short period of time (hours), so they have been replaced performing a linear interpolation process with the closest values. More in details, given two known points of a dataset (x_0, y_0) and (x_1, y_1), the linear interpolant is the straight line between these points. So, a value y of a point (x, y) inside the interval of the two known points is:

$$y = y_0 + (x - x_0)\frac{y_1 - y_0}{x_1 - x_0}$$

A more detailed explanation and a computation of the maximum error that the method can introduce can be read in (22) and in Rolle's Theorem proof (23).

2.1.2 Resampling the dataset for our purpose

The goal of this work is to investigate the possibility to perform a better prediction of 1-day ahead solar radiation divided in 8 samples (3h each) with respect to the forecasting service predictor. Thus, the vector of measurement has been resampled from 15 minutes to 3 hours sample time to be used together with the other datasets. The resampling process consists on

considering just the samples corresponding to the dates *dd/mm/yyyy hh:mm* with:

$$hh = [01; 04; 07; 10; 13; 16; 20; 22];$$
$$mm = [00];$$

discarding then all the others.

Shape of the data

The data is shaped in MATLAB in a matrix that contains as a first column the vector of the dates d in the format expressed in the last paragraph and as a second column the corresponding value x of solar radiation in W/m²:

$$SR_{meas}(t) = \begin{bmatrix} d_t & x_t \\ d_{t+1} & x_{t+1} \\ \vdots & \vdots \\ d_N & x_N \end{bmatrix} \qquad t = 1,2, \dots, N$$

where N is the total number of samples available.

A quick example is shown in Figure 2.

1 Date	2 Value
2014-01-01 01:00:00	0
2014-01-01 04:00:00	0
2014-01-01 07:00:00	0
2014-01-01 10:00:00	264.5000
2014-01-01 13:00:00	509.5000
2014-01-01 16:00:00	37.5000
2014-01-01 19:00:00	0
2014-01-01 22:00:00	0
2014-01-02 01:00:00	0
2014-01-02 04:00:00	0
2014-01-02 07:00:00	0
2014-01-02 10:00:00	9.5000
2014-01-02 13:00:00	36
2014-01-02 16:00:00	16.5000
2014-01-02 19:00:00	0

Figure 2 Example solar radiation matrix

2.2 Solar radiation prediction of the meteorological forecast service (CISMA)

CISMA is a private engineering center that develops environmental models. The predictions are performed by mean of a numerical model called WRF (Weather Research and Forecasting).

Data available

Every 3 hours the prediction till a day ahead is updated. A day-prediction includes 8 predictions (3 hours lagged each) after the date that has been generated.

The predictions available are related to the solar radiation between the 29th June 2016 at 1:00 and 2nd August 2018 1:00. For simplicity in Figure 3 is represented just the data of the first step ahead prediction (3 hours later the generation date).

As it can be easily noticed there are some missing periods, that are:

- From 15th June 2017 10:00 to 29th June 2017 7:00
- From 18th August 2017 19:00 to 6th September 2017 13:00
- From 5th December 2017 19:00 to 12th February 2018 19:00
- From 30th May 2018 10:00 to 20th June 2018 7:00
- From 22th June 2018 22:00 to 28th June 2018 13:00
- From 9th July 2018 10:00 to 10th July 2018 13:00

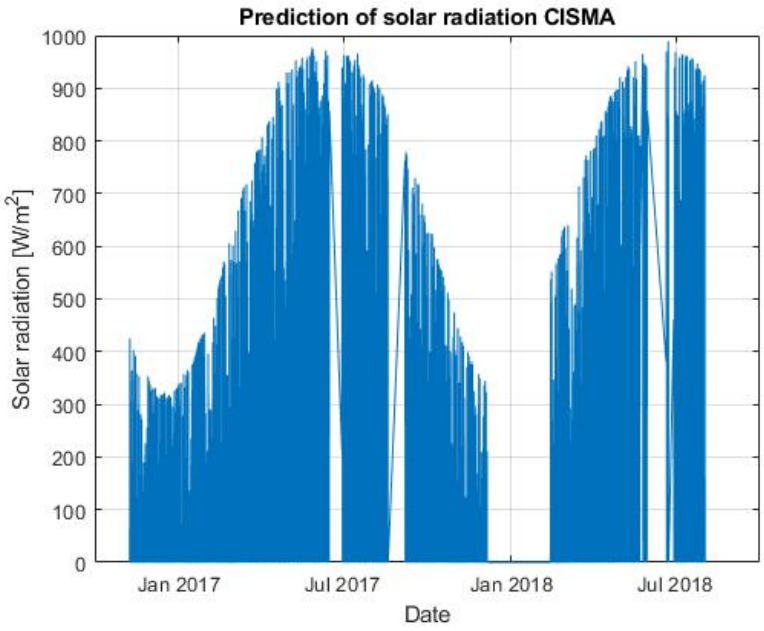

Figure 3 Prediction of solar radiation CISMA

Shape of the data

The data is shaped in MATLAB as a matrix in which the first column is the vector of dates that indicates the time of the first prediction sample (3 hours after the generation), and the remaining columns are the predictions of solar radiation for the whole day after in exactly 8 samples $\widehat{s_t}$ (3h lag each) as depicted below:

$$SR_{CISMA}(t) = \begin{bmatrix} d_t & \hat{s}_{t|t-1} & \hat{s}_{t+1|t-1} & \cdots & \hat{s}_{t+7|t-1} \\ d_{t+1} & s_{t+1|t} & \hat{s}_{t+2|t} & \cdots & \hat{s}_{t+7|t} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ d_M & \hat{s}_{M|M-1} & \hat{s}_{M+1|M-1} & \cdots & \hat{s}_{M+7|M-1} \end{bmatrix} \quad t = 1, \dots, M$$

where M is the number of samples.

The missing periods just described are represented in the matrix by a *Nan* value that differently from the measured solar radiation cannot be replaced with anything, cause the missing intervals are sometimes in the order of

months, so it has been decided to just do not consider those periods in the work.

For clearness an example of the structure of the matrix is shown in Figure 4.

1 Date	2 Value							
08-Nov-2016 07:00:00	0	104	352	55	0	0	0	0
08-Nov-2016 10:00:00	104	352	55	0	0	0	0	0
08-Nov-2016 13:00:00	425	47	0	0	0	0	0	155
08-Nov-2016 16:00:00	47	0	0	0	0	0	155	191
08-Nov-2016 19:00:00	0	0	0	0	0	228	123	1
08-Nov-2016 22:00:00	0	0	0	0	228	123	1	0
09-Nov-2016 01:00:00	0	0	0	218	153	3	0	0
09-Nov-2016 04:00:00	0	0	218	153	3	0	0	0
09-Nov-2016 07:00:00	0	177	165	8	0	0	0	0
09-Nov-2016 10:00:00	177	165	8	0	0	0	0	0
09-Nov-2016 13:00:00	175	10	0	0	0	0	0	117
09-Nov-2016 16:00:00	10	0	0	0	0	0	117	124
09-Nov-2016 19:00:00	0	0	0	0	0	184	364	1
09-Nov-2016 22:00:00	0	0	0	0	184	364	1	0
10-Nov-2016 01:00:00	0	0	0	330	333	10	0	0

Figure 4 Example of Solar radiation prediction CISMA matrix

2.3 Cloud coverage prediction of the meteorological forecast service CISMA

The source of the data and the structure of the predictions are the same depicted in 2.2 so 1-day ahead divided in the whole 8 samples included (3h each). The value predicted consists in cloud coverage expressed in 10 levels. Each level corresponds to a degree of coverage that can have a maximum of 100 (complete covered sky) to a minimum of 0 (clear sky). All levels inside the bounds can assume values in the following vector:

[0 11 22 33 44 56 67 78 89 100]

Data available

The prediction available are between the 8th November 2016 at 07:00 and 2nd August 2018 1:00 with some missing predictions (the same of the solar radiation ones in 2.2). For simplicity in Figure 5 is represented just the data of the first step prediction to let the reader see the shape of the prediction and notice the missing values. As it can be noticed the availability of this data is the smallest.

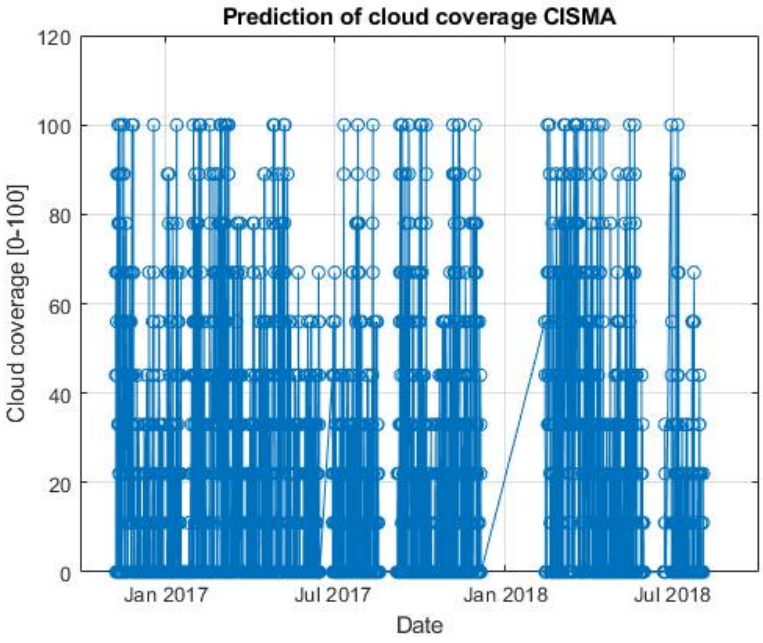

Figure 5 Cloud coverage prediciton CISMA

Shape of the data

The shape of the matrix is the same as the one of solar radiation mentioned in 2.2.

$$CLD_{CISMA}(t) = \begin{bmatrix} d_t & \hat{c}_{t|t-1} & \hat{c}_{t+1|t-1} & \cdots & \hat{c}_{t+7|t-1} \\ d_{t+1} & \hat{c}_{t+1|t} & \hat{c}_{t+2|t} & \cdots & \hat{c}_{t+7|t} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ d_P & \hat{c}_{P|P-1} & \hat{c}_{P+1|P-1} & \cdots & \hat{c}_{P+7|P-1} \end{bmatrix} \quad t = 1, \dots, P$$

where P is the number of samples.

For clearness an example of the structure of the matrix is shown in Figure 6.

1 Date	2 Value							
08-Nov-2016 07:00:00	44	67	33	0	0	0	0	0
08-Nov-2016 10:00:00	67	33	0	0	0	0	0	0
08-Nov-2016 13:00:00	0	0	0	0	0	0	0	44
08-Nov-2016 16:00:00	0	0	0	0	0	0	44	78
08-Nov-2016 19:00:00	0	0	0	0	0	44	67	78
08-Nov-2016 22:00:00	0	0	0	0	44	67	78	67
09-Nov-2016 01:00:00	0	0	0	44	67	78	67	56
09-Nov-2016 04:00:00	0	0	44	67	78	67	56	78
09-Nov-2016 07:00:00	0	56	78	67	89	67	67	44
09-Nov-2016 10:00:00	56	78	67	89	67	67	44	67
09-Nov-2016 13:00:00	67	56	89	56	56	67	67	67
09-Nov-2016 16:00:00	56	89	56	56	67	67	67	67
09-Nov-2016 19:00:00	89	44	67	56	33	44	22	100
09-Nov-2016 22:00:00	44	67	56	33	44	22	100	78
10-Nov-2016 01:00:00	67	0	0	0	22	56	89	100

Figure 6 Example of cloud coverage prediction CISMA matrix

Chapter 3

Statistical models for forecasting

Dealing with the prediction of solar radiation is a hard challenge because, as mentioned in Chapter 1 is strongly dependent to weather conditions.

The process could be studied using very complex mathematical models or by means of statistical models, that are heavily less computationally expensive. A statistical model is a mathematical model that exploits a set of historical data to calibrate its parameters and generate some sample data regarding the future.

The very first step is to choose a model structure able to generate the most precise prediction possible. Thus, in order to find a good structure considering even the computational effort (complexity of the model), it has been chosen to start analyzing first a simple prediction model and then trying to raise the complexity and judge the improvements that they would be able to introduce in terms of prediction error.

Setting calibration and validation datasets

In order to evaluate properly the goodness of a prediction model it is fundamental dividing the available dataset in two parts:

- Calibration dataset
- Validation dataset

The *calibration dataset* has the function to *calibrate* the model *parameters* and the *validation* one is to *test* the same model with the computed *parameters* and build up the *prediction error*.

The choice of the percentage (length) of data used for calibration and validation is very important because, from one side, more data to calibrate the model is dedicated, more the parameters goes to the optimal value, in fact it can be demonstrated that $N \to \infty$, with N number of samples used to calibrate the parameters, leads to the prediction error computed using those parameters to be asimtotically white noise (optimal predictor) (24). From the other side instead, considering that N is finite (we do not have infinite amount of data available), choosing a larger calibration dataset means shorter validation dataset that could lead to having few data to evaluate the model.

In particular, it has been chosen to validate each model with a validation dataset of at least one year, to catch the behavior of the predictors in each season at least once. The final choice has been to dedicate the data from 2th of August 2017 1:00 to 2th of August 2018 1:00 for validation, and the remaining part is used for calibration as shown in Figure 7.

The same validation data set has been used to evaluate the goodness of each tested model.

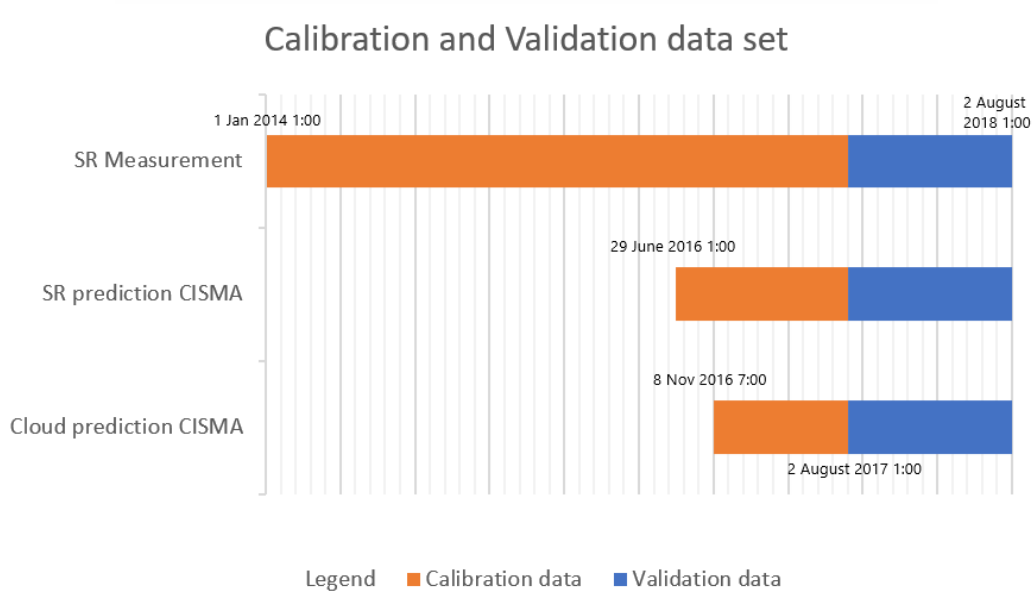

Figure 7 Division Calibration and Validation dataset

3.1 Auto Regressive Integrated Moving Average model (ARIMA)

As a first approach it has been tested the Auto Regressive Integrated Moving Average model (ARIMA).
The ARIMA is a class of models that takes as input a stationary process and returns the forecast of the same process. Any non-seasonal time series that exhibits patterns and is not a random white noise can be modeled with ARIMA models.

It is characterized by 3 terms: p, d, q where p is the order of the Auto Regressive term, q is the order of the Moving average term and d is the number of differencing required to make the eventually non-stationary time series a stationary one.

The form of an ARIMA(p,D,q) is:

$$\Delta^D y_t = c + \phi_1 \Delta^D y_{t-1} + \ldots + \phi_p \Delta^D y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

Where $\Delta^D y_t$ denotes a D^{th} differenced time series, and ε_t is an uncorrelated innovation process with mean zero.

In lag operator notation, $L^i y_t = y_{t-i}$, it can be wrote the ARIMA(p,D,q) model as:

$$\phi^*(L) y_t = \phi(L)(1-L)^D y_t = c + \theta(L)\varepsilon_t$$

where $\phi^*(L)$ is an unstable AR operator polynomial with exactly D unit roots.

You can factor this polynomial as $\phi(L)(1-L)^D$, where $\phi(L) = (1 - \phi_1 L - \ldots - \phi_p L^p)$ is a stable degree p AR lag operator polynomial (with all roots lying outside the unit circle). Similarly, $\theta(L) = (1 + \theta_1 L + \ldots + \theta_q L^q)$ is an invertible degree q MA lag operator polynomial (with all roots lying outside the unit circle). (25)

Stationary process

A time series is weakly stationary if the mean and the variance are constant and finite. Differently, a non-stationary process has time variant mean and variance.

Formally, a stochastic process Y_t is strictly stationary of order n if any n-tuple $(t_1, t_2, \ldots, t_n)$, where k $\in \mathbb{Z}$, the following holds:

$$F_{Y_{t_1},\ldots,Y_{t_n}}(y_{t_1}, \ldots, y_{t_n}) = F_{Y_{t_{1+k}},\ldots,Y_{t_{n+k}}}(y_{t_{1+k}}, \ldots, y_{t_{n+k}}),$$

i.e. if the joint distribution functions of $\{Y_{t_1}, Y_{t_2}, \dots, Y_{t_n}\}$ and $\{Y_{t_{1+k}}, Y_{t_{2+k}}, \dots, Y_{t_{n+k}}\}$ are the same.

For a real-valued process the mean function is defined as $\mu_t = E[Y_t]$ and the variance function is $\sigma_t^2 = E[(Y_t - \mu)^2]$.

A natural estimator of the mean of a single realization of process Y_t $t = 1,2, \dots N$; is:

$$\bar{Y} = \frac{1}{N} \sum_{t=1}^{N} Y_t$$

And the variance of $\bar{Y}$ is:

$$var(\bar{Y}) = \frac{1}{N^2} \sum_{t=1}^{N} \sum_{s=1}^{N} Cov(Y_s, Y_t)$$

where

$$Cov(Y_t, Y_{t+k}) = E[(Y_t - \mu)(Y_{t+k} - \mu)]$$

3.1.1 Deseasonalization process for calibration dataset

As mentioned, the stationarity of the input signal (historical data) is very important to guarantee a good result of the prediction in this type of model.

Thus, considering the strong seasonality shape of the data of solar radiation, a deseasonalization process is needed on the calibration dataset, because otherwise the model would catch the poles at the frequencies of the seasonal terms and not the random process.

The signal can be divided in two parts:

$$x(t) = s(t) + \mu(t)$$

- $s(t) = s(t + kT) \quad k \in Z$ Seasonality (deterministic signal) with period T
- $\mu(t)$ Random part (not deterministic signal)

T in this specific case is easy to catch because of the nature of the data itself. It can be a day, and a year due to the repeating of the seasons.

To understand which is the most relevant T to our purpose (daily predictions), and so to make the signal stationary, the *autocorrelation function* (ACF) and the *partial autocorrelation function* (PACF) have been computed and analyzed for 60 lags. A detailed explanation on how to perform ACF and PACF can be found in (26).

Figure 8 and Figure 9 show respectively the result of ACF and PACF, and it is clear that the data has a strong periodicity of a day, because the ACF and the PACF show peaks at each 8 lags (1 day with 3h sample time).

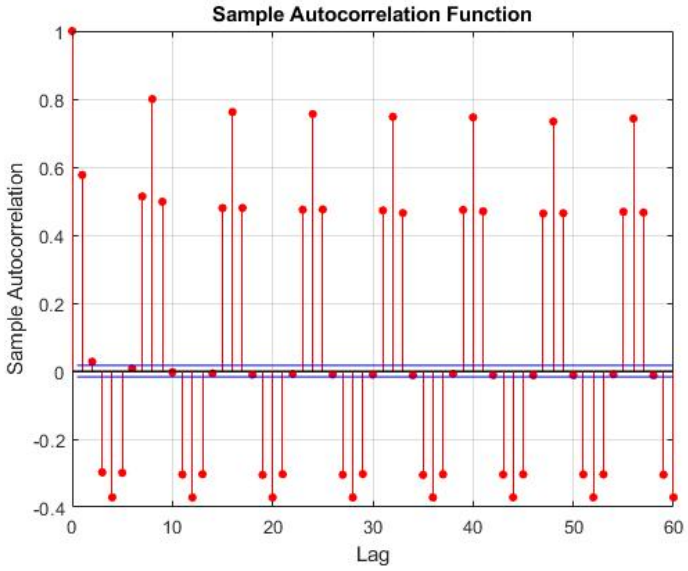

Figure 8 Autocorrelation Function of Solar radiation measurements

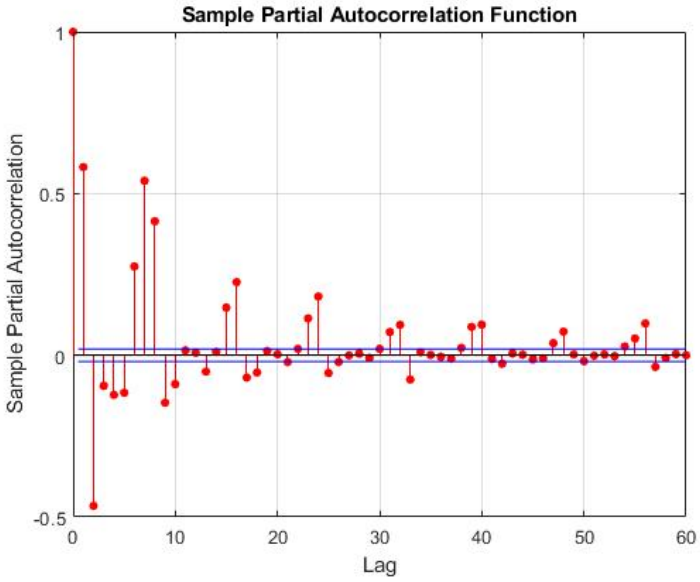

Figure 9 Partial Autocorrelation Function of Solar radiation measurements

Thus, the signal must be made stationary before to be used to calibrate any model.

Seasonality computation by differentiating (ARI)

In order to remove the seasonality from the process, a possible strategy (the most used in literature) is to differentiate the data with the same data at 8 lags behind, it means that each value in the dataset as been differentiated by the value of the instant regarding the day before:

$$\tilde{x}(t) = x(t) - x(t-k) \quad t = 1,2,\dots,N-k$$
$$k = 8; \quad N = number\ samples$$

Figure 10 shows the signal before the differentiation process (figure above) and after (figure below). It is easily noticeable that the yearly periodicity is not removed. To do a more precise analysis the ACF and PACF has been performed again for the differentiated signal $\tilde{x}(t)$, and as it can be noticed in Figure 11 the ACF does not have significant peaks beyond the 8th lag anymore and so, the daily seasonality is to consider removed, while the yearly one is not. But, considering that the prediction horizon in this work is 1-day, yearly seasonality does not affect relevantly the calibration of the parameters because it is considered a slowly-varying process compared with the daily one and so other differentiations are considered unnecessary.

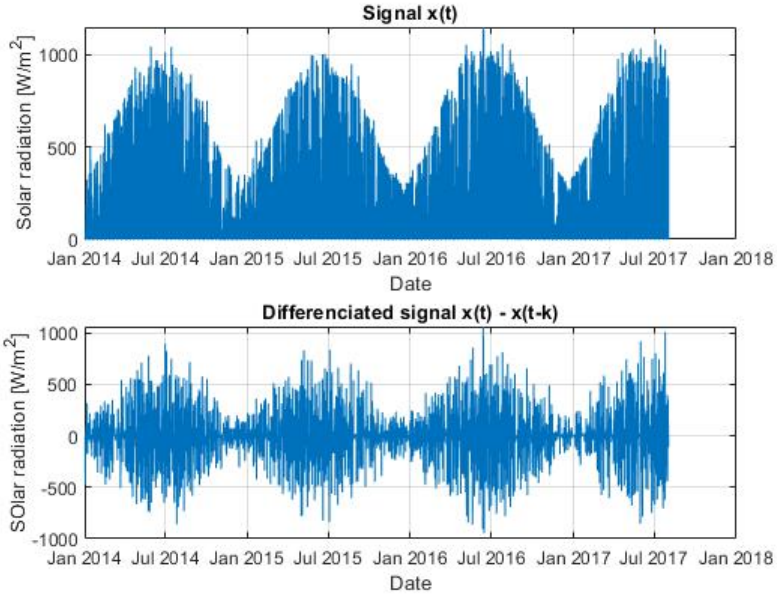

Figure 10 Solar radiation measurements and deseasonalized solar radiation (differentiating)

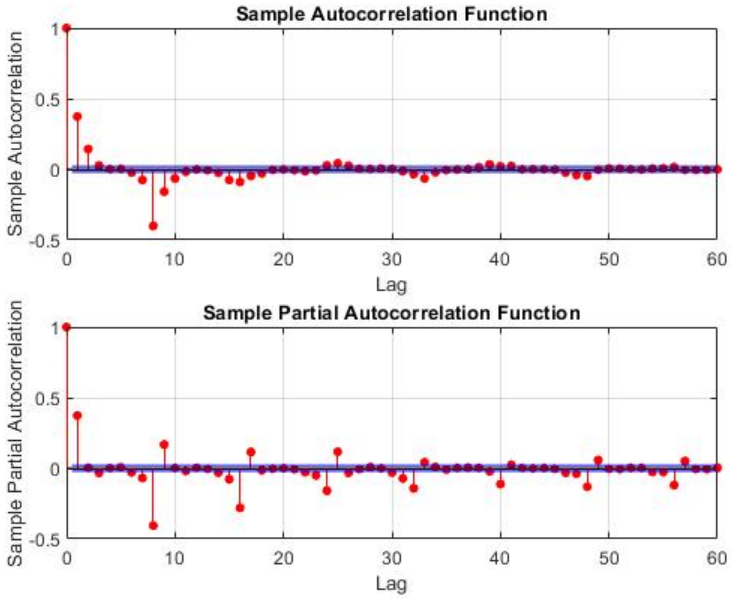

Figure 11 ACF and PACF of Deseasonalized Solar radiation (differentiating)

Seasonality computation by averaging (SAR)

Computing the random signal using the strategy just explained (by differentiating) has a weakness. If the day before the one we must predict was a cloudy one, we basically are removing a seasonality that does not reflect the real typical behavior without the clouds. So, to solve this problem a different approach to deseasonalize the data has been followed too. The objective was to build a data-based signal that reproduce as precisely as possible the deterministic part of the data to be then removed from the complete signal to obtain the random process. To realize it, a good strategy would have been computing for each sample (each 3 hours) the mean of all the data corresponding to the same date. For example, if we want to compute the typical value of the 1st of January at 13:00, we should compute the mean value of all the 1st of January at 13:00 of each year we have available. The problem with this strategy is that it requires a big amount of years of data, because if, for example, we have just 4 years available and 1 or 2 1st of January was cloudy, or rainy, or foggy, the mean is dropping down and the typical value of a sunny day of 1st of January at 13:00 is compromised.

Thus, considering that we have just 4 years available, the approach followed consists on selecting from the calibration dataset the samples belonging to each month and separating them. Then, having the data of each month, the values corresponding to the same hour of the day of that month are averaged forming the mean profile of the month analyzed. The procedure is done for all the 12 months obtaining the typical profile of a year.

$$P_i(k) = \sum_{j=1}^{n_{days_i}} \frac{x_{i,j}(k)}{n_{days_i}}$$

where k is the hour of each day, $x_{i,j}$ is the solar radiation measured of the of the day $j = 1, \dots, n_{days_i}$ of the month $i = 1, \dots, 12$.

Putting the vector found together in a column vector the profile of the year is:

$$P = (P_1 \quad P_2 \quad \cdots \quad P_{12})^T$$

Now that the seasonality is estimated, the random part of the data is obtained removing the seasonality from the original data:

$$\tilde{x}(t) = x(t) - P(t)$$

The Figure 12 shows the original data above and the deseasonalized one below.

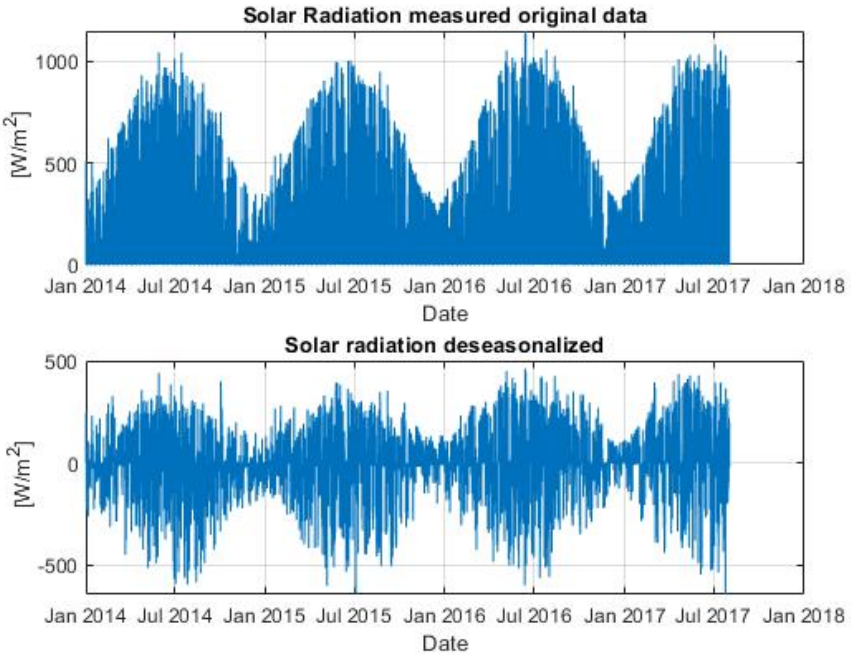

Figure 12 Solar radiation measurements and deseasonalized solar radiation (averaging)

As before ACF and PACF are computed to the signal $\tilde{x}(t) = x(t) - P(t)$ obtaining Figure 13, and for the identical reason just explained the signal is considered stationary and ready to be used to calibrate the models.

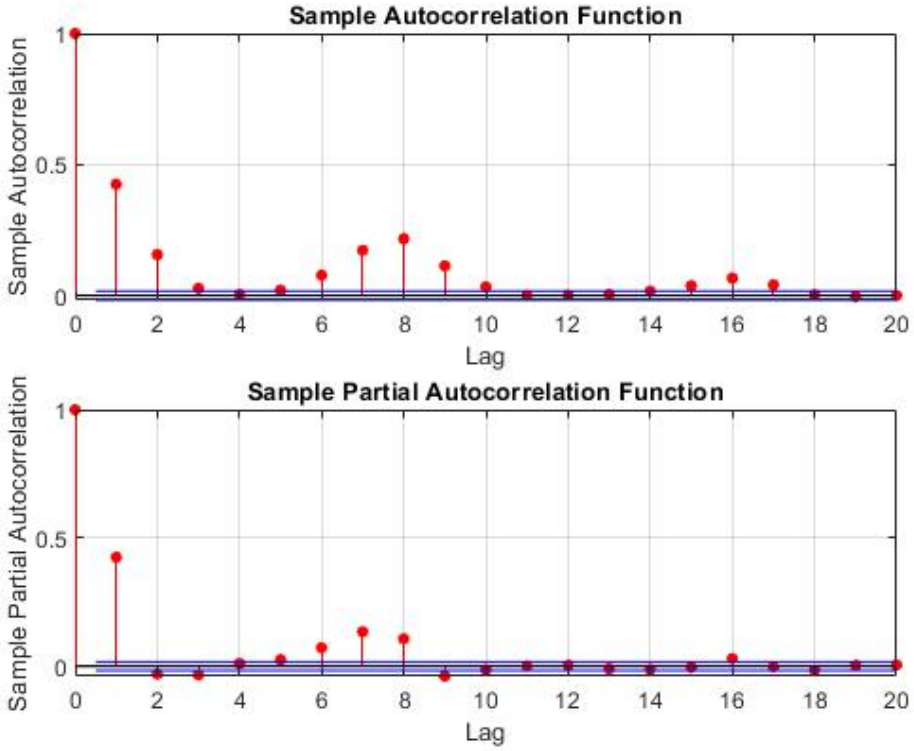

Figure 13 ACF and PACF of Deseasonalized Solar radiation (averaging)

3.1.2 Order choice

After the time series has been made stationary, the next step is to determine whether AR or MA terms are needed to correct any autocorrelation that remains in the differenced or deseasonalized series of solar radiation data.

The literature proposes a systematic way to set the orders of AR and MA terms by looking at ACF and PACF of the differenced time series. The rules used are:

- If the PACF of the differenced series displays a sharp cutoff and/or the lag-1 autocorrelation is positive, then consider adding an AR term to the model. The lag at which the PACF cuts off is the indicated number of AR terms. (27)
- If the ACF of the differenced series displays a sharp cutoff and/or the lag-1 autocorrelation is negative, then consider adding an MA term to the model. The lag at which the ACF cuts off is the indicated number of MA terms. (27)

Considering these rules and taking a look on the Figure 11 and Figure 13 the model suggested is an ARIMA with AR(8) and MA(8).

Actually, after an idea was made through these rules a few simulation has been done and was discovered that the introduction of the MA terms does not introduce any improvement in terms of prediction error, and so the model chosen in order to avoid unnecessary complexity is an ARIMA(8,8,0) with the no MA terms.

3.1.3 Mathematical representation and choice of identification criterion

The mathematical representation of the predictor is then the following:

$$[y(t-h) \quad \cdots \quad y(t-1)] = [y(t-h) \quad \cdots \quad y(t-1)] \begin{bmatrix} a_{11} & \cdots & a_{1h} \\ \vdots & \ddots & \vdots \\ a_{h1} & \cdots & a_{hh} \end{bmatrix}$$

$$t = 1,2,\dots,N-h \qquad h = 8$$

Where $\hat{y}$ is the prediction, y is the data deseasonalized, N the number of samples and h the prediction horizon.

In order to identify the parameters a Prediction Error Minimization (P.E.M) technique has been adopted.

Thus, the following loss function has been used:

$$J_N(\vartheta, k) = \frac{1}{N-h}\sum_{t=h+1}^{N}(y(t) - \hat{y}(t|t-k;\vartheta))^2$$

where $t-k$ indicates the last available sample and $k = 1,2,\dots,8;$

Defining $\varphi(t)^T = \begin{bmatrix} -y(t-h) & \dots & -y(t-1) \\ \vdots & \ddots & \vdots \\ -y(N-h) & \dots & -y(N) \end{bmatrix}$ as the matrix of the data,

$y(t) = \begin{bmatrix} y(t) & \dots & y(t+h-1) \\ \vdots & \ddots & \vdots \\ y(N-h) & \dots & y(N) \end{bmatrix}$ and $\vartheta = \begin{bmatrix} a_{11} & \cdots & a_{h1} \\ \vdots & \ddots & \vdots \\ a_{1h} & \cdots & a_{hh} \end{bmatrix}$ the loss function in a compact notation become:

$$J_N(\vartheta) = \frac{1}{N-h}\sum_{t=h+1}^{N}(y(t) - \varphi(t)^T\vartheta)^2$$

Minimizing the loss function means finding the parameters that bring the derivative of the loss function to zero.

$$\left.\frac{\partial J_N(\vartheta)}{\partial \vartheta}\right|_{\vartheta=\hat{\vartheta}_N} = 0$$

Thus,

$$\frac{\partial J_N(\vartheta)}{\partial \vartheta} = -\frac{2}{N-h}\sum_{t=1+h}^{N-h}(\varphi(t)\,(y(t) - \varphi(t)^T\vartheta)) = 0$$

$$\Rightarrow \left[\sum_{t=1+h}^{N-h} \varphi(t)\varphi(t)^T\right] \vartheta = \left[\sum_{t=1+h}^{N-h} \varphi(t)y(t)\right]$$

The matrix $\sum_{t=1+h}^{N-h} \varphi(t)\varphi(t)^T$ is square and non-singular, so it is invertible. It can be computed the matrix of the parameters:

$$\hat{\vartheta}_N = \left[\sum_{t=1+h}^{N-h} \varphi(t)\varphi(t)^T\right]^{-1} \left[\sum_{t=1+h}^{N-h} \varphi(t)y(t)\right]$$

This method is called Least Squares Identification technique. (28)

3.2 Auto Regressive with eXogenous input (ARX)

Considering the availability of the prediction of solar radiation and cloud coverage provided by the forecasting service CISMA, a more sophisticated model exploiting those data can be developed. The idea is to use a future data inside the ARIMA model as an eXogenous input and see if it helps to improve the accuracy of the prediction. The computation strategy of the parameters is similar to the one of ARIMA discussed in 3.1.3.

3.2.1 Study of stationarity of the eXogenous input

As explained in 2.2 and in 2.3 the shape of the predictions provided by the forecasting service is:

$$\hat{u}_{CISMA}(t) = \begin{bmatrix} d_t & \hat{u}_{t|t-1} & \hat{u}_{t+1|t-1} & \cdots & \hat{u}_{t+7|t-1} \\ d_{t+1} & \hat{u}_{t+1|t} & \hat{s}_{t+2|t} & \cdots & \hat{u}_{t+7|t} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ d_M & \hat{u}_{M|M-1} & \hat{u}_{M+1|M-1} & \cdots & \hat{u}_{M+7|M-1} \end{bmatrix} \quad t = 1, \dots, M$$

with M the number of available samples.

To understand the type of seasonality to be removed the ACF and PACF has been evaluated for solar radiation obtaining Figure 14, and the daily seasonality is evident as expected.

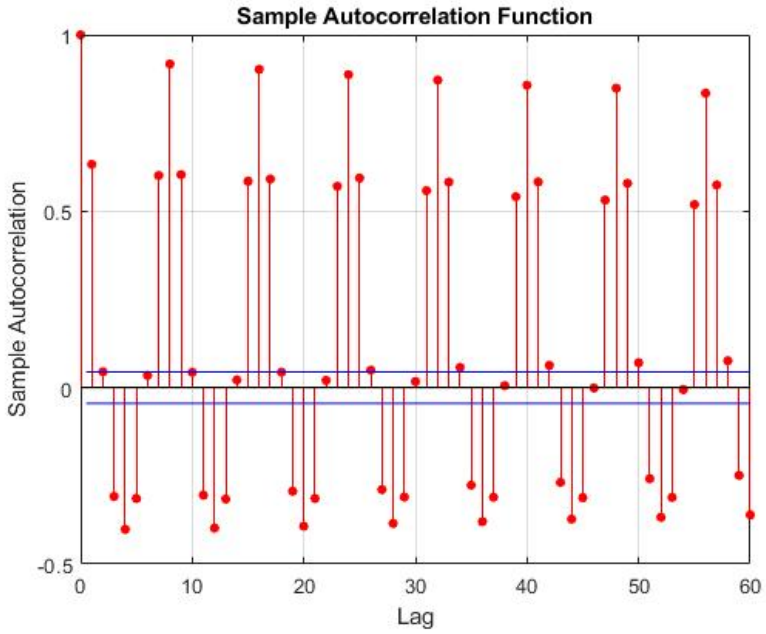

Figure 14 ACF of solar radiation prediciton CISMA

In order to deseasonalize even the eXogenous input the seasonality by averaging computed in 3.1.1 has been removed from each prediction obtaining the deseasonalized eXogenous input.

3.2.2 Order choice

The order of the AR part of this model has been chosen according to the same concepts studied for ARIMA (ACF and PACF plots) in 3.1.2 while for the order of X terms has been decided to exploit every prediction sample available in each instant, so exactly 8 values that results in 8 parameters.

3.2.3 Mathematical representation and choice of identification criterion

The model in a mathematical representation is the following:

$$\begin{bmatrix} \hat{y}(t|t-1) \\ \vdots \\ \hat{y}(t+h-1|t-1) \end{bmatrix}^T = \begin{bmatrix} y(t-h) \\ \vdots \\ y(t-1) \end{bmatrix}^T \begin{bmatrix} a_{11} & \cdots & a_{1h} \\ \vdots & \ddots & \vdots \\ a_{h1} & \cdots & a_{hh} \end{bmatrix} + \begin{bmatrix} \hat{u}(t|t-1) \\ \vdots \\ \hat{u}(t+h-1|t-1) \end{bmatrix}^T \begin{bmatrix} b_{11} & \cdots & b_{1h} \\ \vdots & \ddots & \vdots \\ b_{h1} & \cdots & b_{hh} \end{bmatrix}$$

Where $\hat{y}$ is the prediction, y is the deseasonalized measurement, $\hat{u}$ is the eXogenous input deseasonalized and h the prediction horizon.

The identification technique used is equivalent to the Least Squares described in section 3.1.3.

If it is assumed that the predictor returns the correct future data, it can be formulated the following system of equations:

$$Y = \Phi\vartheta$$

where $$Y = \begin{pmatrix} y(t) & \cdots & y(t+h-1) \\ \vdots & \ddots & \vdots \\ y(N-h) & \cdots & y(N) \end{pmatrix};$$

$$\Phi = \begin{pmatrix} y(t-h) & \cdots & y(t-1) & \hat{u}(t) & \cdots & \hat{u}(t+h-1) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ y(N-2h) & \cdots & y(N-h) & \hat{u}(N-h) & \cdots & \hat{u}(N) \end{pmatrix};$$

$$\vartheta = \begin{pmatrix} a_{11} & \cdots & a_{h1} & b_{11} & \cdots & b_{h1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{1h} & \cdots & a_{hh} & b_{1h} & \cdots & b_{hh} \end{pmatrix}^T$$

The system is overdetermined and the matrix Φ is not invertible because it is rectangular and not squared.

To solve this problem, can be left-multiplied the right and the left side of the equation $Y = \Phi\vartheta$ by Φ^T obtaining $\Phi^T Y = [\Phi^T\Phi]\vartheta$;

The matrix $[\Phi^T\Phi]$ is squared and invertible, thus:

$$\hat{\vartheta} = [\Phi^T \Phi]^{-1} \Phi^T Y$$

It can be proven that the resulting parameter matrix $\hat{\vartheta}$ is the same of the one obtained with the classical Least square's technique. (28)

In this work has been tried multiple variants of this model, one taking as eXogenous input $\hat{u}$ the prediction of solar radiation, another one taking the prediction of cloud coverage, and the last one using both.

3.3 ARX using solar radiation and cloud coverage predictions

Introduction

The amount of solar irradiance that reaches the ground and so the photovoltaic panel is strongly affected by the intermittency of cloud cover. Trivially, the sun is covered by the clouds and so the surface irradiance drops due to attenuation of the direct component. As (8) states though the surface irradiance in some cloud condition can even increase since some clouds tend to enhance diffused radiation.

According to the authors of the document (8) Numerical Weather Prediction Models (NWP) can give accurate predictions in the range of 6 hours to several days. Predictions using satellite images outperform NWP models but in the range from 30 minutes to 6 hours. Thus, in order to be precise in predicting the solar irradiance a good prediction of cloud coverage must be performed. This fact gave the idea to consider the cloud prediction up to 1 day ahead as an input for our model.

Mathematical representation and identification criteria

The mathematical representation varies slightly in the following way:

$$\begin{bmatrix} \hat{y}(t|t-1) \\ \vdots \\ \hat{y}(t+h-1|t-1) \end{bmatrix}^T = \begin{bmatrix} y(t-h) \\ \vdots \\ y(t-1) \end{bmatrix}^T \begin{bmatrix} a_{11} & \cdots & a_{1h} \\ \vdots & \ddots & \vdots \\ a_{h1} & \cdots & a_{hh} \end{bmatrix} + \begin{bmatrix} \hat{u}_{SR}(t|t-1) \\ \vdots \\ \hat{u}_{SR}(t+h-1|t-1) \end{bmatrix}^T \begin{bmatrix} b_{11} & \cdots & b_{1h} \\ \vdots & \ddots & \vdots \\ b_{h1} & \cdots & b_{hh} \end{bmatrix} +$$

$$\begin{bmatrix} \hat{u}_{CLD}(t|t-1) \\ \vdots \\ \hat{u}_{CLD}(t+h-1|t-1) \end{bmatrix}^T \begin{bmatrix} c_{11} & \cdots & c_{1h} \\ \vdots & \ddots & \vdots \\ c_{h1} & \cdots & c_{hh} \end{bmatrix}$$

Where $\hat{y}$ is the prediction, y is the data deseasonalized, $\hat{u}_{SR}$ is the deseasonalized prediciton of solar radiation provided by CISMA, $\hat{u}_{CLD}$ is the prediction of cloud coverage and h the prediction horizon.

$\hat{u}_{CLD}$ has not been deseasonalized because it does not have any correlation with the daily seasonality and so it is stationary by itself.

The identification procedure is the same of the one described in the section 3.2.3 except the shape of the matrices Φ and ϑ, that are in this case:

$$\Phi = \begin{pmatrix} y(t-h) & \cdots & y(t-1) & \hat{u}_{SR}(t) & \cdots & \hat{u}_{SR}(t+h-1) & \hat{u}_{CLD}(t) & \cdots & \hat{u}_{CLD}(t+h-1) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ y(N-2h) & \cdots & y(N-h) & \hat{u}_{SR}(N-h) & \cdots & \hat{u}_{SR}(N) & \hat{u}_{CLD}(N-h) & \cdots & \hat{u}_{CLD}(N) \end{pmatrix}$$

$$\vartheta = \begin{pmatrix} a_{11} & \cdots & a_{1h} & b_{11} & \cdots & b_{1h} & c_{11} & \cdots & c_{1h} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{h1} & \cdots & a_{hh} & b_{h1} & \cdots & b_{hh} & c_{h1} & \cdots & c_{hh} \end{pmatrix}^T$$

Where N is the number of samples available for all the data at the same time, h is the prediction horizon, $\hat{u}_{SR}$ is the solar radiation prediction of CISMA, and $\hat{u}_{CLD}$ the cloud coverage prediction of CISMA.

And thus, the parameter estimated is given by:

$$\hat{\vartheta} = [\Phi^T \Phi]^{-1} \Phi^T Y$$

Chapter 4

Ensemble methods

Introduction

Different types of forecasting models have been tested in the research, each type, in most of the cases, is able to capture different aspects of the information available on the data for the prediction with respect to another. For this reason, can be worth exploiting the benefits that each model can offer. Thus, researchers are constantly working on this field and Ensemble forecasting is one of relatively modern method discovered. According to Fengxia Zheng and Shouming Zhong in (29), forecasting accuracy can be improved through the combination of multiple individual forecasts. In (11) the ensemble development consists on averaging the results of two predictions, while in (29) a linear combination of the predictions is performed. As F. Zheng and S. Zhong in (29) state, the weight of each objective is computed minimizing the distance from the measurements.

In this chapter similar methods have been tested and will be discussed. The schematic representation of the averaging method ensemble is illustrated in Figure 15, and the one of the linear combination one in Figure 16. They are for simplicity represented as the combination of just two models, but the combination could be extended easily to n models.

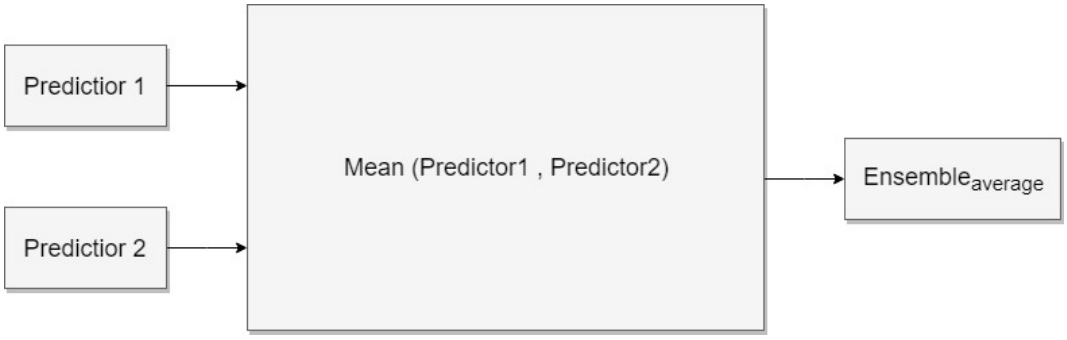

Figure 15 Ensemble method representation (averaging)

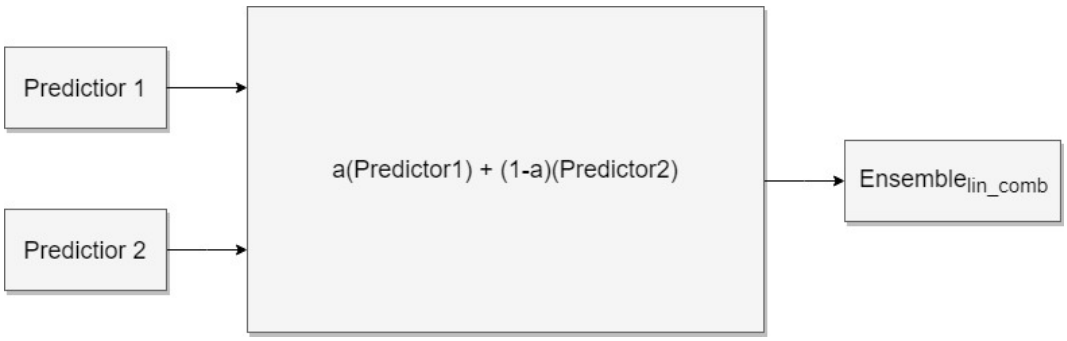

Figure 16 Ensemble method representation (linear combination)

4.1 Ensemble using the prediction obtained by the model ARIMA in 3.1 and solar radiation prediction (CISMA)

Averaging method

The first ensemble ES1 consists on averaging the results of the Auto Regressive model discussed in 3.1 and the prediction of solar radiation provided by the forecasting service in 2.2.

To do so, the vector of Solar radiation measurements has been reshaped in the following format (the same of the format of the output of each predictor):

$$Y = \begin{pmatrix} y(t) & \cdots & y(t+h) \\ \vdots & \ddots & \vdots \\ y(N-h) & \cdots & y(N) \end{pmatrix}$$

where N is the number of the predicted samples and h is the prediction horizon. The equation of the ensemble ES1 is then the following:

$$ES1(i,j) = \left(\frac{\hat{Y}_{AR}(i,j) + \hat{Y}_{CISMA}(i,j)}{2}\right)_{\substack{i=1,..,N \\ j=1,...,8}}$$

Where $\hat{Y}_{AR}(i,j)$ is the prediction obtained by the predictor ARIMA of section 3.1 and $\hat{Y}_{CISMA}(i,j)$ the prediction of solar radiation of CISMA.

Linear combination method

The idea of the linear combination of the two predictors is a mathematical way to understand how much faith each predictor deserves with respect to another one. So, the second ensemble ES2 is computed with the idea to find a parameter α that fulfill as precisely as possible the following equation:

$$\alpha\hat{Y}_{AR} + (1-\alpha)\hat{Y}_{CISMA} = Y$$

$$\hat{Y}_{CISMA} = \begin{pmatrix} \hat{Y}_{CISMA}(1,1) & \cdots & \hat{Y}_{CISMA}(1,h) \\ \vdots & \ddots & \vdots \\ \hat{Y}_{CISMA}(N,1) & \cdots & \hat{Y}_{CISMA}(N,h) \end{pmatrix},$$

$$\hat{Y}_{AR} = \begin{pmatrix} \hat{Y}_{AR}(1,1) & \cdots & \hat{Y}_{AR}(1,h) \\ \vdots & \ddots & \vdots \\ \hat{Y}_{AR}(N,1) & \cdots & \hat{Y}_{AR}(N,h) \end{pmatrix}.$$

$$Y = \begin{pmatrix} y(t) & \cdots & y(t+h) \\ \vdots & \ddots & \vdots \\ y(N-h) & \cdots & y(N) \end{pmatrix}$$

where N is the number of the predicted samples and h is the prediction horizon.

In order to compute the parameter α all the data have been divided in calibration and validation part and the same procedure as the equivalent of Least Squares explained previously has been adopted, thus:

$$\hat{\alpha} = [\Phi^T \Phi]^{-1} \Phi^T \Psi$$

where $\Phi = \hat{Y}_{AR_{cal}} - \hat{Y}_{CISMA_{cal}}$ and $\Psi = Y_{cal} - \hat{Y}_{CISMA_{cal}}$

The parameter has been calibrated using the calibration dataset and validated with the remaining part in the following way:

$$ES2 = \hat{\alpha}\, \hat{Y}_{AR_{val}} + (1 - \hat{\alpha}) \hat{Y}_{CISMA_{val}}$$

The result is finally compared with Y_{val} and the error evaluated.

4.2 Ensemble using the prediction obtained by the model ARX in 3.2 and solar radiation prediction (CISMA)

The ensembles ES3 and ES4 are computed with the exact same idea of the one in 4.1 but while the predictor in input for ES1 and ES2 was $\hat{Y}_{AR}$, in this one is $\hat{Y}_{ARX}$, which is the result of the predictor proposed in the paragraph 3.3.

Thus, the predictors are:

$$ES3(i,j) = \left(\frac{\hat{Y}_{ARX}(i,j) + \hat{Y}_{CISMA}(i,j)}{2} \right)_{\substack{i=1,..,N \\ j=1,...,8}}$$

with *averaging method*

and the following for the *linear combination* approach as 4.1:

$$ES4 = \hat{\beta}\,\hat{Y}_{ARX_{val}} + \left(1 - \hat{\beta}\right)\hat{Y}_{CISMA_{val}}$$

with $\hat{\beta}$ computed in the following way:

$$\hat{\beta} = [\Phi^T \Phi]^{-1} \Phi^T \Psi$$

where $\Phi = \hat{Y}_{ARX_{cal}} - \hat{Y}_{CISMA_{cal}}$ and $\Psi = Y_{cal} - \hat{Y}_{CISMA_{cal}}$

where the subscript "cal" means the calibration part of the data.

The result is compared with Y_{val} as before and the error evaluated, where the subscript "val" means the validation part of the data.

Chapter 5

Auto Regressive with eXogenous input with labels concept

The basic idea of Ensemble methods as said, is to try to exploit the benefits of each model combining more than one method. The idea of using labels for the prediction is a smarter way to exploit the characteristics of each model used, because the label concept basically refers to find a rule that is able to judge the characteristic of a prediction, and exploit it to use the model that best behave in the found situation. In this chapter, two strategies are investigated. In 5.1 the predictions are labeled as "Good day" and "Bad Day" in terms of accuracy of the prediction, while in 5.2 the goal is to predict the state of the day after within two possible states: Sunny day and Cloudy day.

In 5.3 another version of 5.2 is proposed, which uses just the predictions of cloud coverage to set the label that at this time is divided within three states: Sunny day, Medium clear sky day and Cloudy day.

5.1 Labels Good and Bad prediction day

The goodness of the ARX models discussed in the previous chapters has a strong dependency on the goodness of the prediction made by the forecasting

service (CISMA). Thus, a possible idea was to understand when and why the predictor of CISMA makes a wrong prediction and try to develop a model that sets the parameters based on that. To do so, it has been initially analyzed the prediction error of the solar radiation predictions of CISMA, and then its possible correlation with the solar radiation and cloud coverage predictions was investigated. The idea of introducing *labels* consist on judging the prediction of each day by CISMA comparing it with the measurement of solar radiation of the same day and labeling it as a "*Good prediction day*" or a "*Bad prediction day*" based on the criteria explained in the next section. Then, using the label data just generated, calibrate a Support Vector Machine (SVM) model to let it predict the label of the day after given the one of the day before. After that, the idea is to use the predicted label to switch from a dedicated model (GDARX – Good Day ARX) to the other (BDARX Bad Day ARX).

5.1.1 Labels Good and Bad prediction predicted by a Support Vector Machine (SVM)

Basic theoretical concepts on Support Vector Machine for Binary Classification

The literature (30) states that if we must predict a data that has exactly two classes the SVM for binary classification can be used. An SVM classifies data by finding the best hyperplane that separates all data points of one class from those of the other class. The *best* hyperplane for an SVM means the one with the largest *margin* between the two classes. *Margin* means the maximal width of the slab parallel to the hyperplane that has no interior

data points.

The *support vectors* are the data points that are closest to the separating hyperplane; these points are on the boundary of the slab. The Figure 17 illustrates these definitions, with + indicating data points of type 1, and – indicating data points of type –1.

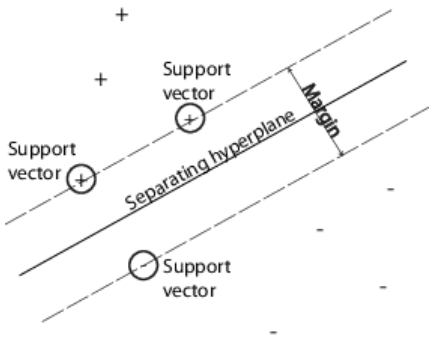

Figure 17 Support Vector Machine representation

The data for training is a set of points (vectors) x_j along with their categories y_j. For some dimension d, the $x_j \in R^d$, and the $y_j = \pm 1$. The equation of a hyperplane is:

$$f(x) = x'\beta + b = 0$$

The following problem defines the *best* separating hyperplane (i.e., the decision boundary). Find β and b that minimize $||\beta||$ such that for all data points (x_j, y_j)

$$y_j f(x_j) \geq 1$$

The support vectors are the x_j on the boundary, those for which

$$y_j f(x_j) = 1$$

A more detailed theoretical concepts of SVM can be seen in (30).

Label estimation procedure

In order to develop the vector of labels briefly explained in the last paragraph the following procedure was followed and implemented in MATLAB environment.

The first step was to define a rule to judge each day predicted by the forecasting service CISMA. Considering that the prediction consists on predicting 1 day ahead every 3h as explained before, the Normalized Root Mean Square Error of each sample of the prediction has been computed.

Initially the data of measurements x_{SR} are reshaped in the way to let them fit with the shape of the prediction $\hat{u}_{SR}$, so:

$$Y = \begin{pmatrix} y(t) & \cdots & y(t+h) \\ \vdots & \ddots & \vdots \\ y(N-h) & \cdots & y(N) \end{pmatrix}$$

and

$$\hat{u}_{SR} = \begin{pmatrix} \hat{u}_{SR}(t) & \cdots & \hat{u}_{SR}(t+h) \\ \vdots & \ddots & \vdots \\ \hat{u}_{SR}(N-h) & \cdots & \hat{u}_{SR}(N) \end{pmatrix}$$

where N is the total number of samples and h the prediction horizon.

The NRMSE of each sample (3h) is computed as follows:

$$err_{pred_{3h}}(i) = \frac{\sqrt{\sum_{j=1}^{8} \frac{[y(i,j) - \hat{u}_{SR}(i,j)]^2}{8}}}{\max\{y(i,j)\}_{\substack{i=1,\dots N \\ j=1,\dots,8}}}$$

But, whereas the goal is to judge as said the goodness of the prediction of 1 day of CISMA and not each 3h, the mean of the values of $err_{pred_{3h}}$ corresponding to the same day of the year has been computed and it generates the vector $err_{pred_{1day}}$ containing the mean of the Normalized Root Mean Square Error of each day.

$$err_{pred_{1day}}(k_{day}) = \left(\frac{\sum err_{pred_{3h}}(j(k_{day}))}{n_{samples_{k_{day}}}}\right)$$

where k_{day} is the k[th] day , $j(k_{day})$ the indexes of the errors that belongs to the k[th] day and $n_{samples_{k_{day}}}$ is the amount of $j(k_{day})$ that each error has.

The rule to set a label for each value of the just found vector respects the following criteria:

$$if \left|err_{pred_{1day}}(k_{day})\right| \geq \varepsilon \;\Rightarrow\; Label(k_{day}) = 0 \text{ ("}Bad\ day\text{ ")}$$
$$else \;\Rightarrow\; Label(k_{day}) = 1 \text{ ("}Good\ day\text{")}$$
$$where\ \varepsilon = 20;$$

The value of the threshold ε is arbitrary, it has been chosen in the way that the algorithm would not consider too many days as "Good" compared with "Bad" ones (it would mean that it is too permissive), and too many "Bad" days with respect to "Good" ones (it would mean too pretentious).

The idea now, is to calibrate an SVM that is able, as precisely as possible, to predict the label of a day given the label of the day before and the mean of the prediction of cloud coverage for the day after. It has been tried to calibrate the SVM using also the prediction of solar radiation of the forecasting service CISMA, but it returns a less precise result, so it was not considered later.

So, in order to be put in the calibration vectors of the SVM, the vector of mean cloud coverage for each day $CLD_{mean_{1day}}$ has been computed.

$$CLD_{mean_{3h}}(i) = \sum_{j=1}^{8} \frac{[\hat{u}_{CLD}(i,j)]}{8}$$

and

$$CLD_{mean_{1day}}(k_{day}) = \left(\frac{\sum CLD_{mean_{3h}}(j(k_{day}))}{n_{samples_{k_{day}}}}\right)$$

where k_{day} is the k[th] day , $j(k_{day})$ the indexes of the cloud values that belongs to the k[th] day and $n_{samples_{k_{day}}}$ is the amount of $j(k_{day})$ that each cloud value has.

Now, all the vector needed to calibrate and validate the label's predictor (SVM) are available.

To do so, the usual calibration and validation division is made ($Label_{cal}, CLD_{mean_{1day_{cal}}}$,$Label_{val}$ and $CLD_{mean_{1day_{val}}}$) for both of vector and the matrices Y and X are defined in the following way:

$$X_{cal} = \begin{pmatrix} Label_{cal}(k_{day}) & CLD_{mean_{1day_{cal}}}\left(k_{day_{cal}} + 1\right) \\ \vdots & \vdots \\ Label_{cal}(N_{cal} - 2) & CLD_{mean_{1day_{cal}}}(N_{cal} - 1) \end{pmatrix};$$

$$Y_{cal} = \begin{pmatrix} Label_{cal}(k_{day} + 1) \\ \vdots \\ Label_{cal}(N_{cal} - 1) \end{pmatrix};$$

with N_{cal} is the number of calibration samples and $k_{day} = 1,2 \ldots, N_{cal}$.

X and Y are used to train the SVM model by means of the function in MATLAB *fitcsvm(X_{cal}, Y_{cal}).*

Once calibrated the SVM model, the validation data set X_{val} computed with the same criteria as the calibration one is used to generate the vector of predicted labels $label_{pred}$.

$$X_{val} = \begin{pmatrix} Label_{val}(k_{day}) & CLD_{mean_{1day_{val}}}\left(k_{day} + 1\right) \\ \vdots & \vdots \\ Label_{val}(N_{val} - 2) & CLD_{mean_{1day_{val}}}(N_{val} - 1) \end{pmatrix}$$

with N_{val} is the number of validation samples and $k_{day} = 1,2,\dots,N_{val}$.

The vector $label_{pred}$ is computed in MATLAB as follows:

$$label_{pred} = predict(SVM_{model}, X_{val})$$

Finally, each label predicted is compared respectively with the observed one and the error is evaluated in the following way:

$$\widehat{err}\% = \frac{n^0(diff\{label_{SVM}(k) - label(k)\} \neq 0)}{N_{val}} 100 = 14\%;$$

$$k = 1,2,\dots,N_{val};$$

5.1.2 ARX models depending on labels (GDARX – BDARX)

The models to be calibrated are with the same identical structure of the ARX of the section 3.2.3 but they have dedicated parameters depending on the label as Figure 18 illustrates.

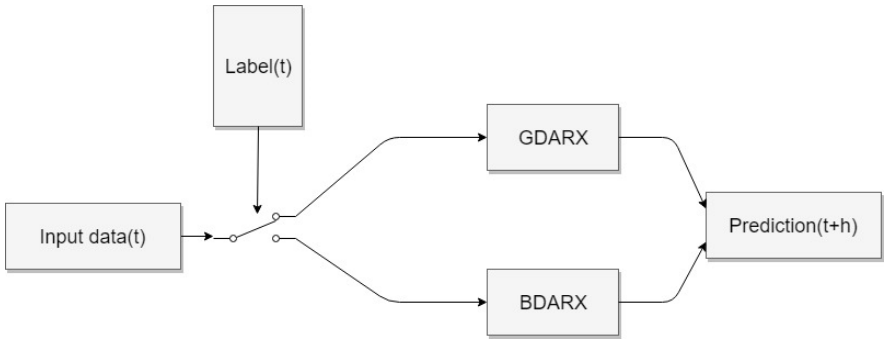

Figure 18 GB-ARX representation

The mathematical representation of GD-ARX and BD-ARX then is:

$$\begin{bmatrix} \hat{y}(t|t-1) \\ \vdots \\ \hat{y}(t+h-1|t-1) \end{bmatrix}^T = \begin{bmatrix} y(t-h) \\ \vdots \\ y(t-1) \end{bmatrix}^T \begin{pmatrix} a_{11} & \cdots & a_{1h} \\ \vdots & \ddots & \vdots \\ a_{h1} & \cdots & a_{hh} \end{pmatrix} + \begin{bmatrix} \hat{u}_{SR}(t|t-1) \\ \vdots \\ \hat{u}_{SR}(t+h-1|t-1) \end{bmatrix}^T \begin{pmatrix} b_{11} & \cdots & b_{1h} \\ \vdots & \ddots & \vdots \\ b_{h1} & \cdots & b_{hh} \end{pmatrix} +$$

$$\begin{bmatrix} \hat{u}_{CLD}(t|t-1) \\ \vdots \\ \hat{u}_{CLD}(t+h-1|t-1) \end{bmatrix}^T \begin{pmatrix} c_{11} & \cdots & c_{1h} \\ \vdots & \ddots & \vdots \\ c_{h1} & \cdots & c_{hh} \end{pmatrix} \qquad t = 1,2,\dots,N-h;$$

where N is the number of samples to predict and h is the prediction horizon (8 samples, 1 day), $\hat{u}_{SR}$ and $\hat{u}_{CLD}$ respectively the prediction of solar radiation of CISMA and the prediction of cloud coverage of CISMA.

Calibration of the models

To calibrate the models the same procedure of Least Squares is followed but different calibration datasets are used. For the model GD-ARX the calibration dataset is composed by days with "Good day" label, and for BD-ARX just by days with "Bad day" label.

Doing this, we obtain two matrices of parameters $\hat{\vartheta}_{GD}$ and $\hat{\vartheta}_{BD}$.

$$\hat{\vartheta}_{GD} = \begin{pmatrix} a_{GD_{11}} & \cdots & a_{GD_{1h}} & b_{GD_{11}} & \cdots & b_{GD_{1h}} & c_{GD_{11}} & \cdots & c_{GD_{1h}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{GD_{h1}} & \cdots & a_{GD_{hh}} & b_{GD_{h1}} & \cdots & b_{GD_{hh}} & c_{GD_{h1}} & \cdots & c_{GD_{hh}} \end{pmatrix}^T$$

$$\hat{\vartheta}_{BD} = \begin{pmatrix} a_{BD_{11}} & \cdots & a_{BD_{1h}} & b_{BD_{11}} & \cdots & b_{BD_{1h}} & c_{BD_{11}} & \cdots & c_{BD_{1h}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{BD_{h1}} & \cdots & a_{BD_{hh}} & b_{BD_{h1}} & \cdots & b_{BD_{hh}} & c_{BD_{h1}} & \cdots & c_{BD_{hh}} \end{pmatrix}^T$$

Once the label is predicted by the SVM model in each prediction instant the parameter to use for that specific prediction is chosen following the criteria:

$$if\ Label_{pred}\big(K_{day}(t)\big) = 0 \quad \Rightarrow \quad \hat{\vartheta} = \hat{\vartheta}_{BD}$$
$$else\ if\ Label_{pred}\big(K_{day}(t)\big) = 1 \quad \Rightarrow \quad \hat{\vartheta} = \hat{\vartheta}_{GD}$$

where t the date to predict and $K_{day}(t)$ is the day corresponding to the date t.

5.2 Label Sunny-Cloudy prediction by SVM method

Once evaluated the result of the prediction in 5.1.2, has been noticed that the predictor makes wrong predictions more likely when it does not recognize perfectly if the day to predict will be sunny, medium clear sky or cloudy. Thus, the idea was to use a similar system to predict the label of the next prediction with the algorithm of Support Vector Machine with two classes. The label can assume this time the following two states:

- Sunny day
- Cloudy day

Then once the label is predicted, it will be used to add the deterministic part of the signal (the one studied in 3.1.1) multiplied by a constant parameter that model a typical Sunny day, or a typical Cloudy day of the corresponding period of the year.

Label estimation procedure

As a first step a classifier for each day of the measured solar radiation has been developed. It evaluates for each sample its label. The classifier takes as input a day and it compares its maximum value of solar radiation with the maximum value of the corresponding typical day.

Using a threshold ∂ evaluates the label in the following way:

$$if\ \mathrm{SRmax}\{day\} \geq \partial\ \mathrm{SRmax}\{day_{typical}\} \quad \Rightarrow \quad Label(day) = 1\ (Sunny)$$
$$else\ if\ \ \mathrm{SRmax}\{day\} < \partial\ \mathrm{SRmax}\{day_{typical}\} \ \Rightarrow \quad Label(day) = 0\ (Cloudy)$$

Where ∂ has been set to 0.7.

Using this criteria, the labels for solar radiation measurements are computed and we call it $Label_Y(t)$ with $t = 1, \dots, N$.

$Label_Y(t)$ is divided in calibration part (to calibrate the SVM) and validation part (to test the SVM) with the same length of the ones to calibrate and validate the models.

The SVM model is calibrated using the following calibration matrices:

$$X_{cal} = \begin{pmatrix} Label_{Ycal}(t) & CLD_{mean_{1day_{cal}}}(t+h) \\ \vdots & \vdots \\ Label_{Ycal}(N_{cal}-h) & CLD_{mean_{1day_{cal}}}(N_{cal}) \end{pmatrix};$$

$$Y_{cal} = \begin{pmatrix} Label_{Ycal}(t+h) \\ \vdots \\ Label_{Ycal}(N_{cal}) \end{pmatrix};$$

Where N_{cal} is the number of samples of calibration h the prediction horizon and $CLD_{mean_{1day_{cal}}}$ the same as 5.1.1.

So, as in 5.1.1 using the MATLAB function *fitcsvm(X_{cal}, Y_{cal})* the SVM is calibrated.

Once calibrated the SVM model, the validation dataset X_{val} is used to obtain the predicted labels $label_{pred}$ in the following way:

$$label_{pred} = predict(SVM_{model}, X_{val})$$

$$X_{val} = \begin{pmatrix} Label_{val}(k_{day}) & CLD_{mean_{1day_{val}}}(k_{day}+1) \\ \vdots & \vdots \\ Label_{val}(N_{val}-2) & CLD_{mean_{1day_{val}}}(N_{val}-1) \end{pmatrix}$$

with N_{val} is the number of validation samples and $k_{day} = 1,2, \dots, N_{val}$.

Finally, each label predicted is compared respectively with the observed one and the error is evaluated in the following way:

$$\widehat{err}\% = \frac{n^0(diff\{label_{SVM}(k) - label(k)\} \neq 0)}{N_{val}} 100 = 23\%;$$

$$k = 1,2, \dots, N_{val};$$

ARX model depending on labels Sunny-Cloudy (SCARX)

As explained in the introduction of the paragraph, the label here is used to choose to add a typical Sunny day or a typical Cloudy day of that period to the prediction, using the model ARX in 3.3 with the same parameters. The typical Cloudy day has been chosen as 70% of a Sunny one.

5.3 Label Sunny-Medium-Cloudy prediction by cloud coverage predictions

The solution in 5.2 (with the classes "Sunny day" and "Cloudy day") has the limitation of having just two possible states, and this imply to do not model properly every state in the middle of Sunny one and Cloudy one. To solve partially that limitation a solution with three possible states has been tried.

In this model the label can be: Sunny day, Medium clear sky day and Cloudy day and it is predicted just evaluating the mean value of the predictions of cloud coverage of the day after. To do it, the following rule has been set:

$$if\ \mathrm{CLD}_{\mathrm{mean}}\{k_{day}\} \geq \partial_{max} \quad \Rightarrow \quad Label\{k_{day}\} = 1\ (Sunny)$$
$$else\ if\ \ \mathrm{CLD}_{\mathrm{mean}}\{k_{day}\} < \partial_{min} \quad \Rightarrow \quad Label\{k_{day}\} = -1\ (Cloudy)$$
$$else\ if \quad \partial_{min} \leq \mathrm{CLD}_{\mathrm{mean}}\{k_{day}\} < \partial_{max} \ \Rightarrow \quad Label(day) = 0\ (Medium)$$

With $\partial_{max} = 70$ and $\partial_{min} = 30$

The error in predicting the label is computed as:

$$\widehat{err}\% = \frac{n^0(diff\{label_{CLD}(k) - label(k)\} \neq 0)}{N_{val}} 100 = 28{,}99\%;$$

$$k = 1,2, \dots, N_{val};$$

ARX model depending on labels Sunny-Medium-Cloudy (SMCARX)

The model used to perform the prediction is the ARX in 3.3 with the identical parameters. A similar procedure of 5.2 to add the typical profile is used, but this time having three states has been computed the typical profile for each possible state, that in particular it resulted to multiply the typical Sunny profile to 0.5, 0.8 and 1 respectively for Cloudy day, Medium clear sky day and Sunny day.

Chapter 6

Results Analysis

6.1 Prediction Error Analysis

Introduction

Today, there is a debate regarding the best metrics for evaluating the accuracy of a forecast in the industry. Some researches have been done for this purpose as said in (31) (32) (33) (34), but as (35) says "at present there is no industry-wide standard performance metric". The metric to evaluate the prediction error of each model depends on the purpose for which is developed. Several well accepted metrics are summarized in (35). The common indicators though are Root Mean Square Error (RMSE), Mean Percentage Error (MPE), Mean Absolute Percentage Error (MAPE) and Mean Bias Error (MBE). Each of them can catch different characteristics of a model. Really important is also studying the distribution of the prediction error, that means evaluating its mean value and its variance.

The Mean Percentage Error (MPE) indicates the average ratio of deviations to the actual values while the Mean Absolute Percentage Error (MAPE) simply takes the absolute value of the MPE. The Mean Biased Error (MBE) provides information on the long-term performance, over/under-estimation of the model in the long run. The Root Mean Square Error instead, due to its

definition may read a high value even if only a single measurement has high deviation from its model generated counterpart.

In this precise case it has been used a normalized version of the RMSE (NRMSE).

It has been decided to use just the $NRMSE_{\%}$, MPE and MAPE because the analysis of this thesis is a short/medium-term one and so MBE is not considered.

The definitions of the errors are in general:

$$MPE = \frac{1}{N_{val}} \sum_{i+1}^{N_{val}} \left(\frac{\hat{Y}_i - Y_i}{Y_i} \right);$$

$$MAPE = \frac{1}{N_{val}} \sum_{i+1}^{N_{val}} \left| \frac{\hat{Y}_i - Y_i}{Y_i} \right|;$$

$$NRMSE = \frac{\sqrt{\frac{1}{N_{val}} \sum_{i+1}^{N_{val}} \left(\hat{Y}_i - Y_i \right)^2}}{\max(Y_1, \dots, Y_{N_{val}})}$$

Considering though that the prediction horizon is 8 lags in this case and so each prediction is formed by 8 values, the metrics to judge the whole prediction are particular and are explained below.

Normalized Root Mean Square Error for 1 day prediction horizon

The RMSE is computed for each prediction separately, obtaining a vector of RMSEs and then the $NRMSE_{tot}\%$ is computed as the mean of the RMSEs computed for each prediction.

$$RMSE_{k_{pred}}(k_{pred}) = \sqrt{\frac{1}{8}\sum_{i=1}^{8}\left(\hat{Y}_i(k_{pred}) - Y_i(k_{pred})\right)^2}$$

where k_{pred} is the k-th prediction, i represents the lag-ahead of the prediction, $\hat{Y}_i$ is the predicted value and finally Y_i is the measured value.

$$NRMSE_{tot}\% = \frac{\frac{1}{N_{val}}\sum_{i=1}^{N_{val}}\left(RMSE_{k_{pred}}(i)\right)}{\max(Y_1, \dots, Y_{N_{val}})}100$$

where N_{val} is the total number of the predictions.

Mean Percentage Error (MPE) and Mean Absolut Percentage Error (MAPE)

The same approach is followed so the MPE and $MAPE$ are computed for each prediction separately and their mean is calculated obtaining MPE_{tot} and $MAPE_{tot}$.

$$MPE_{k_{pred}}(k_{pred}) = \frac{1}{8}\sum_{i=1}^{8}\left(\frac{\hat{Y}_i(k_{pred}) - Y_i(k_{pred})}{Y_i(k_{pred})}\right);$$

$$MPE_{tot} = \frac{1}{N_{val}}\sum_{i=1}^{N_{val}}\left(MPE_{k_{pred}}(i)\right);$$

$$MAPE_{k_{pred}}(k_{pred}) = \frac{1}{8}\sum_{i=1}^{8}\left|\frac{\hat{Y}_i(k_{pred}) - Y_i(k_{pred})}{Y_i(k_{pred})}\right|;$$

$$MAPE_{tot} = \frac{1}{N_{val}}\sum_{i=1}^{N_{val}}\left(MAPE_{k_{pred}}(i)\right);$$

Forecast Error Distribution

As mentioned, studying the statistical properties of a prediction permits also to evaluate a predictor. In this section are presented some methods for estimating a confidence interval around a forecast.

One characteristic that a predictor should guarantee is that the forecast $\hat{Y}_i(\tau)$ is *unbiased*. It means that it must satisfy the following equation:

$$E\left[\hat{Y}_t(\tau)\right] = E[Y_t(\tau)]$$

where $Y_t(\tau)$ is the time series.

This imply that $E\left[\hat{Y}_t(\tau) - Y_t(\tau)\right] = 0$ so the mean of the prediction error $\hat{\varepsilon}_\tau = \hat{Y}_t(\tau) - Y_t(\tau)$ should be near zero.

Furthermore, in the case the predictor is fitted correctly it should apply that:

$$\hat{\varepsilon}_\tau = \hat{Y}_t(\tau) - E[Y_t(\tau)] - \eta$$

where η is *white noise* by assumption and so an unpredictable process.

Said this, if $E\left[\hat{Y}_t(\tau) - Y_t(\tau)\right]$ and η are uncorrelated then:

$$Var[\hat{\varepsilon}_\tau] = Var\left[E\left[\hat{Y}_t(\tau) - Y_t(\tau)\right]\right] + Var[\eta] = \sigma_E^2(\tau) + \sigma_\eta^2$$

And therefore, the variance of the error in estimating the future value $Y_t(\tau)$ is the sum of two different variances, the variance caused by the estimation of the mean $\sigma_E^2(\tau)$ and the one of the noise σ_η^2.

Considering that the statistical model was not calibrated with $N \to \infty$ samples the parameters are not perfectly correct and so the model is not perfectly correct as well, this imply that $\hat{\varepsilon}_\tau$ could be not *white noise*.

It is obvious that more the prediction error is close to be white noise, more the predictor is precise.

The sample standard deviation can be evaluated given N samples of forecast errors $\{\varepsilon_i\}_{i=1,..,N}$, and it is given by:

$$\delta_\varepsilon = \sqrt{\frac{\sum_{i=1}^{N}(\varepsilon_i - \bar{\varepsilon})^2}{N-p}}$$

Where $\bar{\varepsilon}$ is the sample average of the error and p the number of parameters.

The value of $\widehat{\delta_\varepsilon}$ is the estimation of the variance of σ_ε^2 from the data, and it includes $\sigma_E^2(\tau)$ and σ_η^2.

So, it is assumed for example that the noise comes from a normal distribution, a confidence interval of 95% estimation of the forecast can be approximated by:

$$\hat{Y}_t(\tau) \pm 1.96\, \widehat{\delta_\varepsilon}$$

Because the prediction horizon is 8 samples ahead and so 1 day in this work $\widehat{\delta_\varepsilon}$ will be the mean of the deviations of each time step ahead prediction.

6.2 Evaluation of the models and comparisons

As discussed in 6.1 section, the NRMSE, MPE, MAPE and a statistical analysis is performed to each model described in the thesis and the results have been compared each other. Further, some example situations are showed to explain the pros and the cons introduced of each model.

6.2.1 Errors generated

In Table 2 and Table 3 can be seen the computed errors explained in 6.1 for each model. The notation used is explained in Table 1.

Model Notation	Model
AR	Auto Regressive Integrated Moving Average (ARIMA) in 3.1 Differentiating
SAR	Auto Regressive Integrated Moving Average (ARIMA) in 3.1 with Mean Seasonality
ARX1	Auto Regressive with eXogenous Input (ARX) using SR_{CISMA} in 3.2
ARX2	Auto Regressive with eXogenous Input (ARX) using CLD_{CISMA} in 3.2
ARX3	Auto Regressive with eXogenous Input (ARX) using $SR_{CISMA,}$ CLD_{CISMA} in 3.3
GB-ARX	ARX models depending on labels (GDARX – BDARX) in 5.1.2
SC-ARX	ARX model depending on labels Sunny-Cloudy **Error! Reference source not found.** in 5.2
SMC-ARX	ARX model depending on labels Sunny-Medium-Cloudy in 5.3
ES1	Ensemble Averaging Method using AR and SR_{CISMA} in 4.1
ES3	Ensemble Averaging Method using ARX3 and SR_{CISMA} in 4.2
ES2	Ensemble Linear Combination Method using AR and SR_{CISMA} in 4.1
ES4	Ensemble Linear Combination Method using ARX3 and SR_{CISMA} in 4.1

Table 1 Notation of the models

Model Name	$NRMSE_{model}$	MPE_{model}	$MAPE_{model}$	$\widehat{\delta_{\varepsilon_{model}}}$	$NRMSE_{CISMA}$	MPE_{CISMA}	$MAPE_{CISMA}$	$\widehat{\delta_{\varepsilon_{CISMA}}}$
AR	10.3493	2.6681	3.0783	139.4476	10.0521	0.9083	1.3700	138.1636
SAR	9.5298	0.7855	1.3309	122.1070	10.0521	0.9083	1.3700	138.1636
ARX1	8.8227	1.0220	1.4588	117.0115	10.0521	0.9083	1.3700	138.1636
ARX2	9.8074	0.9427	1.5213	128.8778	10.0521	0.9083	1.3700	138.1636
ARX3	8.5977	0.5539	1.2068	114.2489	10.0521	0.9083	1.3700	138.1636
GB-ARX	8.5326	2.1709	2.6356	115.4249	10.0521	0.9083	1.3700	138.1636
SC-ARX	8.6574	0.7057	1.2946	119.3020	10.0521	0.9083	1.3700	138.1636
SMC-ARX	8.7357	0.6022	1.2578	120.8326	10.0521	0.9083	1.3700	138.1636
ES1	8.8778	1.7882	2.1140	119.8642	10.0521	0.9083	1.3700	138.1636
ES3	8.9365	0.7290	1.2368	122.7573	10.0521	0.9083	1.3700	138.1636

Table 2 Errors of each model validation 1

Model Name	$NRMSE_{model}$	MPE_{model}	$MAPE_{model}$	$\widehat{\delta_{\varepsilon_{model}}}$	$NRMSE_{CISMA}$	MPE_{CISMA}	$MAPE_{CISMA}$	$\widehat{\delta_{\varepsilon_{CISMA}}}$
ES2	8.8801	1.6725	2.0140	129.9167	10.0518	0.7481	1.1616	140.0572
ES4	8.6891	0.3713	1.1396	128.7043	10.0518	0.7481	1.1616	140.0572

Table 3 Errors of each model validation 2

6.2.2 Results comments

The results highlighted in Table 2 and Table 3 are separated cause the ones in Table 3 are computed using a slightly shorter validation dataset due to the need to use part of the predictions made by the predictors used for the ensembles to calibrate the related parameter for the linear combination.

Interpreting the meaning of the errors, some intuitive comments can be done. Each predictor presented, apart AR, works better than the one of the forecasting service CISMA in terms of NRMSE, MPE, MAPE and standard deviation $\widehat{\delta_{\varepsilon}}$.

Comparing the models in terms of NRMSE, the best one results to be the GB-ARX one but apparently it is the worst after the AR in terms of MPE and MAPE. These considerations suggest that the model GBARX does not make extremely wrong predictions as the ones with higher NRMSE because of the label system that predicts when the prediction will be "Bad" and tries to adjust it, but it is less precise when the others work relatively well cause sometimes the label results to be wrong, this is stated cause the NRMSE is really sensitive to relatively big errors considering that it makes the square of the prediction error, so strong errors are highlighted. In MPE and MAPE instead strong errors are weighted as the lighters and so they suggest that a prediction that have smaller MPE and MAPE behave better in general.

In terms of MPE and MAPE the best models, considering the ones in Table 2 first, are in descending order ARX3, SMC-ARX and ES3, while the model GB-ARX, ES1 and AR are ones that have it the highest. In particular, GB-ARX has a higher MPE and MAPE because in the days in which the predicted label is wrong the model predicts with the wrong parameters that makes the error increase.

In terms of standard deviation $\widehat{\delta_\varepsilon}$ the best six models in descending order are: ARX3, GBARX, ARX1, SCARX, ES1 and SMCARX and every model is better than CISMA.

Considering these intuitive considerations, the model that best fit with the data is ARX3, and the improvements introduced with respect to CISMA are highlighter in Table 4.

Improvement%	I_{NRMSE}%	I_{MPE}%	I_{MAPE}%	$I\delta_{\varepsilon}$%
ARX3/CISMA	14.4686%	39.0179%	11.9124	17.3090

Table 4 Table of improvements % of the best method ARX3 compared with CISMA

The average improvement introduced is 20.677 %, so, a great improvement.

6.2.3 Prediction error analysis and comparison between the most interesting models

In this subsection some statistical analysis is performed to the prediction errors highlighting the main properties and some pros and cons introduced by the different approaches followed.

AR vs SAR

As depicted in Table 2 SAR, with respect to AR, introduce strong improvements on each error analyzed. The structure of the model is the same while the only difference is on deseasonalization process and on re-adding the seasonality on the prediction. AR, as expected, makes strong errors when the weather from a day to another changes, for example if the day before was cloudy, AR most likely predict the day after as a cloudy one, it works well instead if the weather remains similar from a day to the following. In Figure 19 an example of the different behavior of the two model is shown. In the 24th of October of 2017 SAR can track better the measurements than AR, the weather evidently changes from the day before to the day after and AR add wrongly the profile of the day before, which was cloudy while the day after was not. The same happens in the 22th of October of 2017, but this time the

day before was sunny while the day after was cloudy, AR then, overestimates the solar radiation of the day after while SAR does it less. In Figure 20 instead the ACF of both prediction errors of 1 day-lag is shown, both are pretty much white noise, but the ACF of the prediction error of AR presents a negative peak at 8^{th} and 16^{th} lag, so there is a small correlation, this means that the model is not optimum.

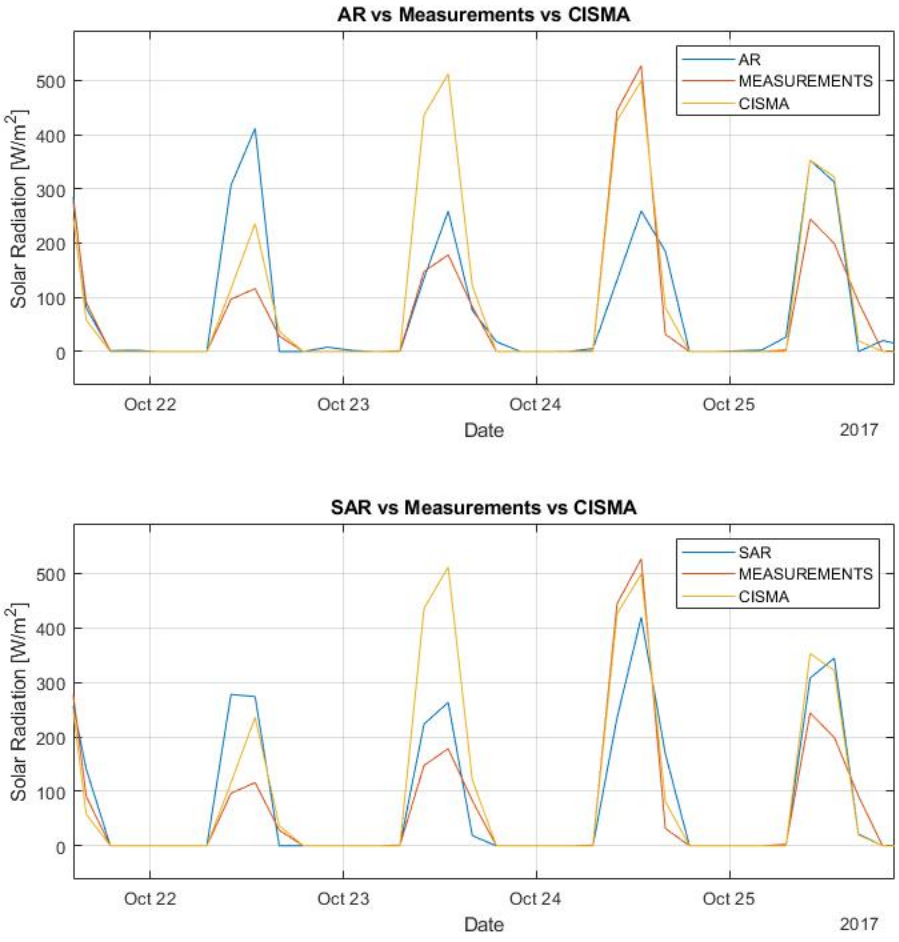

Figure 19 Example of predictions AR vs SAR for comparison

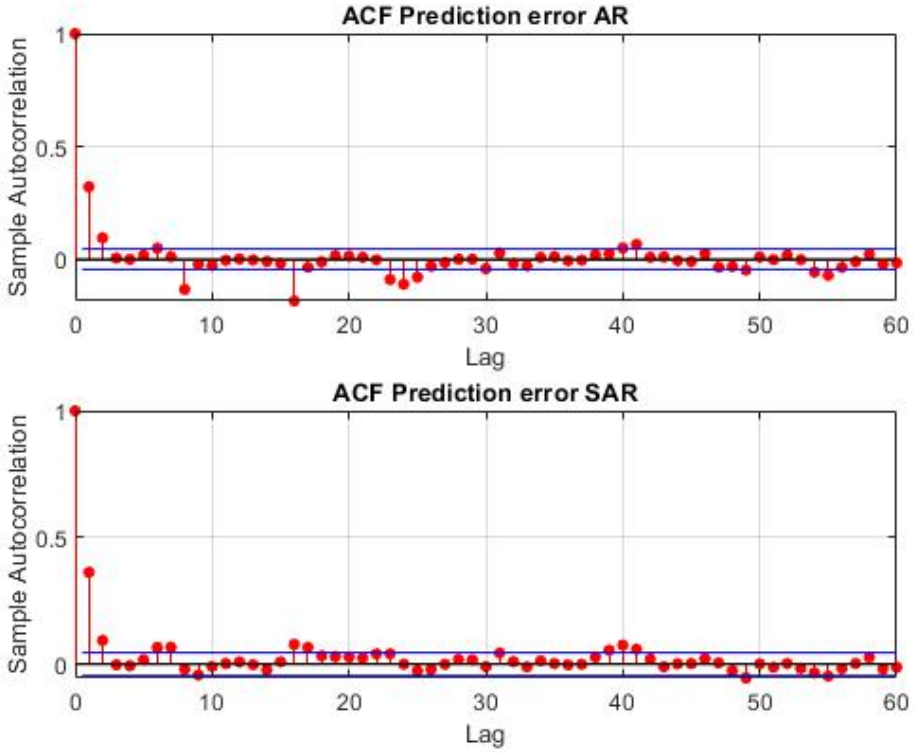

Figure 20 ACF of prediciton error of AR vs SAR for comparison

ARX1 vs ARX2 vs ARX3

In these models the eXogenous inputs are involved, as explained before ARX1 just exploit the predictions of CISMA of solar radiation, ARX2 just the ones of cloud coverage and ARX3 both. In the example of Figure 21 are highlighted the different behavior of the models. It can be noticed that ARX2 with respect to the others has the weakness that if the cloud prediction is wrong, the prediction error increases a lot, while the others are less dependent upon it, cause depends also on the predictions of solar radiation. The 26th of October of 2017 the weather was medium clear sky while the cloud prediction says it is clear sky, in fact the prediction of ARX2 overestimates the solar radiation. The day after is a cloudy day and again the prediction of cloud coverage says

medium clear sky and makes the ARX2 overestimate again. ARX3 exploiting both future data can track better than all the others the solar radiation measurements.

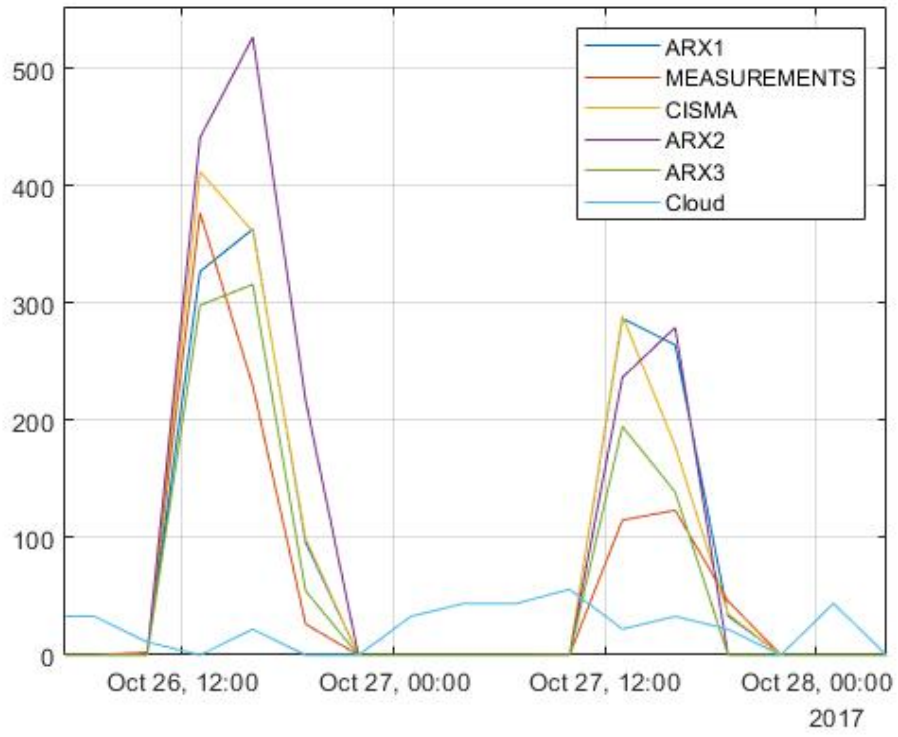

Figure 21 Example of predictions ARX1 vs ARX2 vs ARX3 vs CISMA for comparison

SMC-ARX vs SC-ARX

These two models share the idea to predict the label of cloudiness of the day to be predicted, we remind that SC-ARX divide the label state of each day in Sunny and Cloudy and use a SVM model to predict them, while SMC-ARX divide the label in Sunny, Medium, Cloudy day and does not use a SVM to predict them, but just the mean value of the prediction of cloud coverage.

Standing to the results of NRMSE, MPE and MAPE, the SC-ARX exceeds the quality of SMC-ARX in term of NRMSE but does not for both the other errors.

This suggests that dividing the label in three levels leads to slightly improve the accuracy when the prediction of the label of the corresponding day is right, but when the label is wrong makes higher errors.

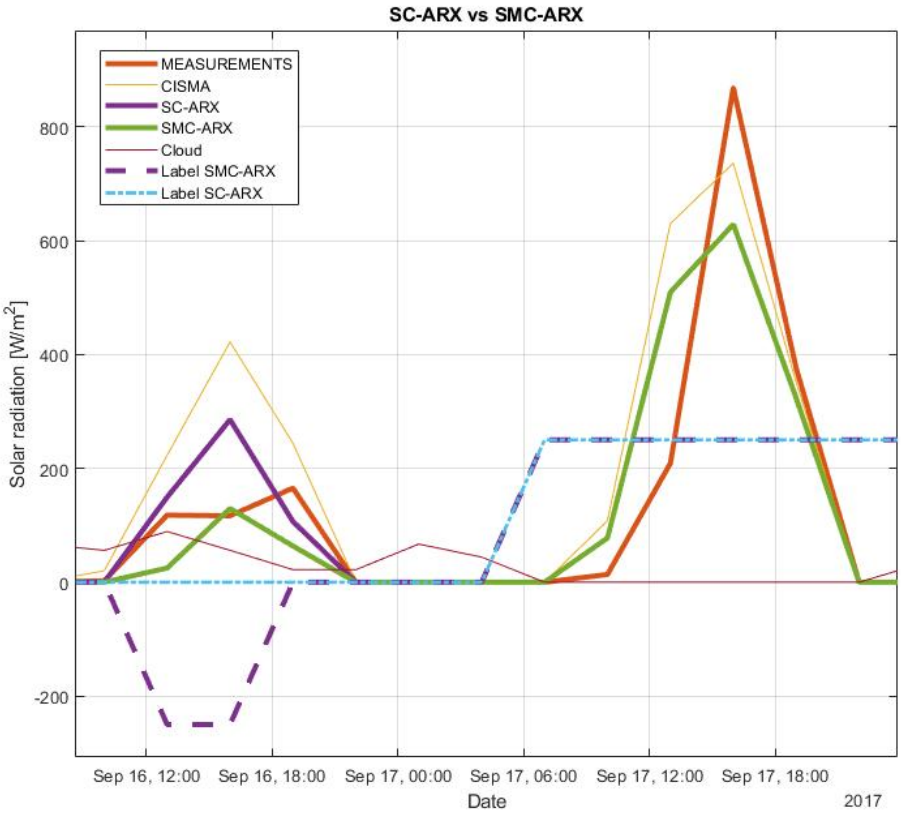

Figure 22 Example of predictions SC-ARX vs SMC-ARX for comparison

In Figure 22 can be seen that in the 16th of September 2017 for example the predictor of the label of SC-ARX does not have the ability to predict that the day is really cloudy while SMCARX does cause the third level. It helps to predict that the solar radiation in that situation will be smaller. In the day after, both SC-ARX and SMC-ARX predict that the day will be a sunny one

and the results of the predictors are overlapped cause the structure and the parameters of the ARX are the same. At the same time the predictor of the label of SMC-ARX has more probability to make a wrong prediction on the label (about 30% of error against 23% of SC-ARX one), and if the choice is wrong the error of the prediction will be higher than the one of SC-ARX. This behavior is shown in the errors computed as well.

The ACFs of the prediction errors depicted in Figure 23 shows that there are some peaks on each multiple of 8^{th} lag for three days, this is probably due to a recurrent error provoked by the label prediction system.

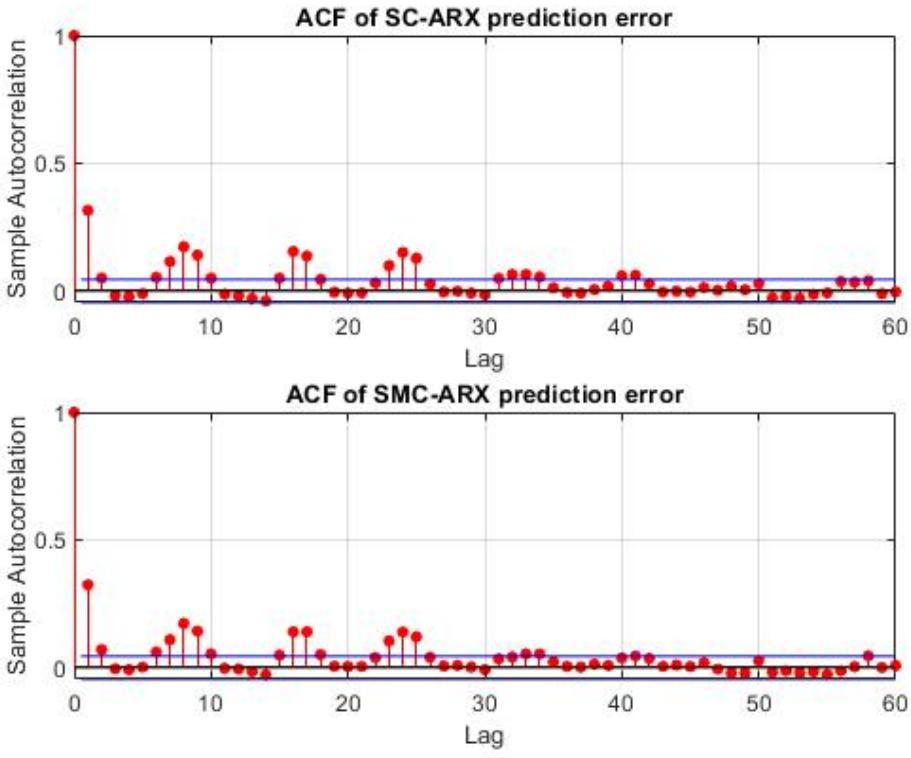

Figure 23 ACF of predictions errors SC-ARX vs SMC-ARX for comparison

ARX3 vs GB-ARX

It has been decided to compare in particular these two models because even though ARX3 has an higher NRMSE it also has very lower MPE and MAPE, this can be attributed at the fact that when the prediction of the label "Good day prediction" or "Bad day prediction" is correct a dedicated model is used that responds in the proper way while ARX3 does not.

In Figure 24 is shown a case in which the prediction of the label for that day is correct and it is "Bad", it can be noticed that while ARX3 overestimates the solar radiation as CISMA does, the model GB-ARX tracks well the solar radiation and makes a lighter error.

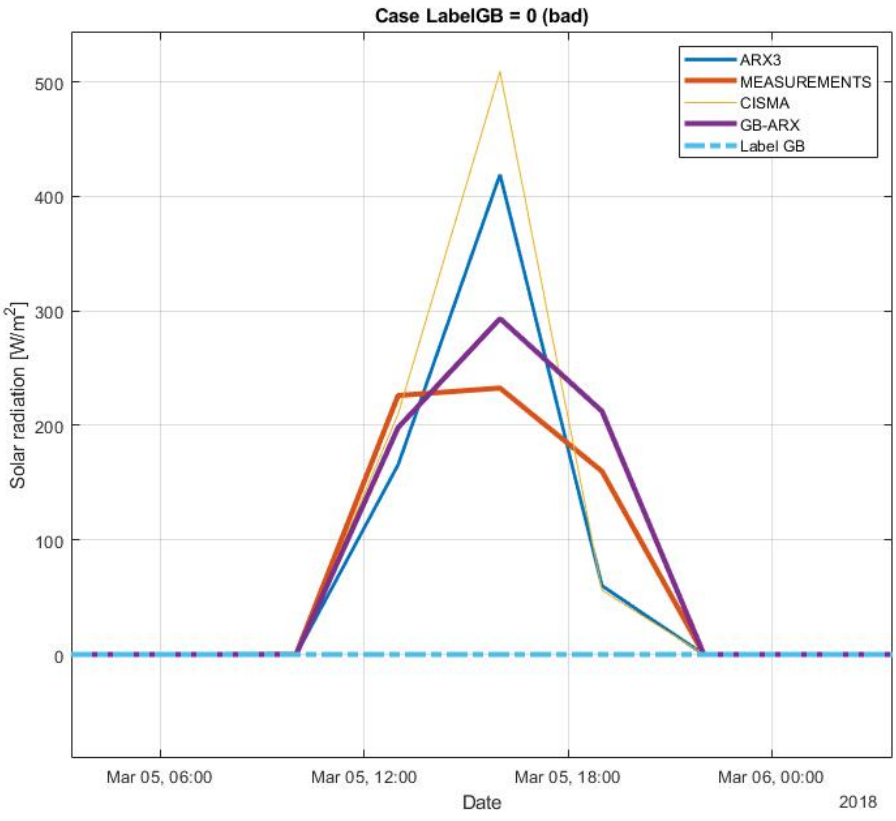

Figure 24 Example of predictions ARX3 vs GB-ARX vs CISMA for comparison

In Figure 25 instead, the dual case in which the prediction of the label for that day is “Bad” while it should not is shown. It can be noticed that when the prediction of label is wrong the model GB-ARX for that day has less accuracy with respect to ARX3.

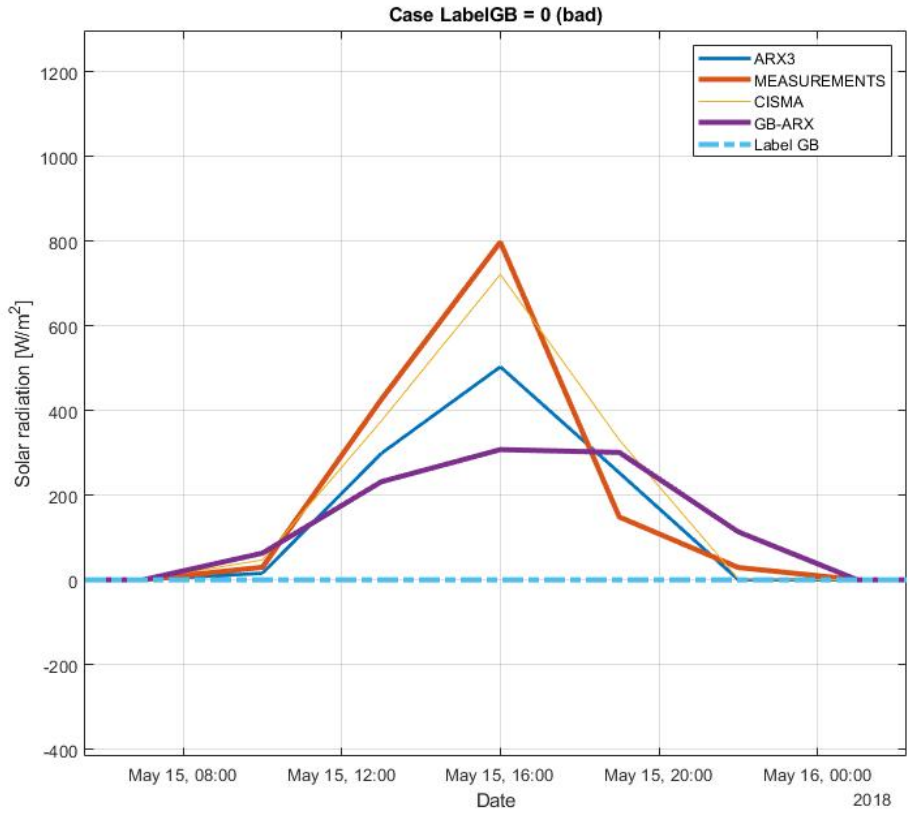

Figure 25 Example of predictions ARX3 vs GB-ARX vs CISMA for comparison

The analysis of ACF of both the prediction errors of both models (as shown in Figure 26) leads to the consideration that the prediction errors are white noise and so the model are optimum.

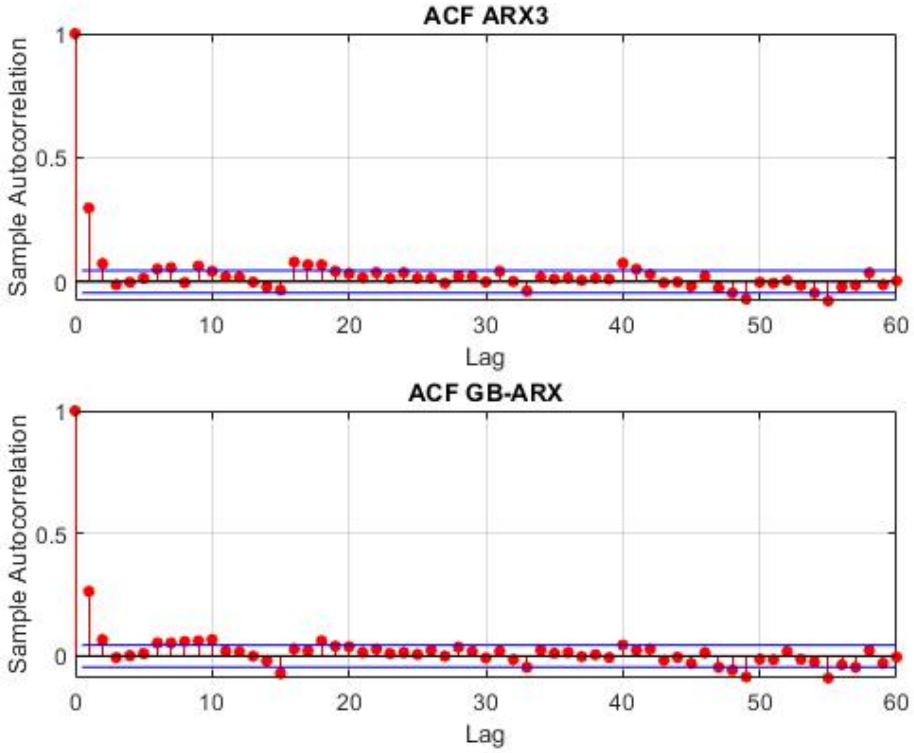

Figure 26 ACF of predictions errors of ARX3 vs GB-ARX for comparison

Chapter 7

Conclusions

As said in the introduction of the thesis, exactly in the background studies in 1.2, pushing in the direction of predicting the solar radiation with a day-ahead prediction horizon using data-driven models as ARIMA combined with complex predictors as CISMA (ARX1, ARX2, ARX3 and Ensembles) and some hybrid model such as ARIMA combined with machine learning methods (SVM) bring better accuracy in the predictions.

In particular, each model has some pros and some cons. Every model presented in the thesis outperforms the one of the forecasting service CISMA except for AR. Some model work even better than ARX3, which is evaluated as the best one, but just in some particular situations. For example SC-ARX and SMC-ARX solve partially the problem of predicting a sunny or cloudy day adding the right typical profile, but they makes higher error in the days in which the prediction of the label are wrong, and considering that the percentage of error of the label for SC-ARX is 23,21% and the one of SMC-ARX is 28,99% the improvement with respect to ARX3 are marginal but still consistent for CISMA ones. Regarding GB-ARX, it exceeds the accuracy of the predictions of ARX3 when the prediction of CISMA of solar radiation (which is based on) is very inaccurate, but at the same time introduces bigger errors in the days in which the label predicted is wrong. The percentage of error for the labels here is 14,63%. Thus, in the

same way it does not introduce several improvements to ARX3 but still outperforms CISMA. The Ensemble methods bring all pretty good results, in particular the ES4, which is a linear combination of ARX3 and CISMA in this way $ES4 = \hat{\beta}\,\hat{Y}_{ARX} + (1-\hat{\beta})\hat{Y}_{CISMA}$ but in general still not better then ARX3.

ARX3 works well almost under every aspect and so it has been considered finally the best one with a percentage of improvement of 20.677% with respect to CISMA. (which is the mean of the improvements% in Table 4).

Future Works

It has been noticed the potentiality of combining very different typologies of models. Considering what explained in 6.2.3, the models have different qualities. A good way to proceed could be investigating on a particular Ensemble method that recognizes as precise as possible the situations to use a model rather than another. Furthermore, it could be worth to investigate to some more methods to predict the labels with a higher accuracy permitting the introduced models to work in its own ideal situation. Thus, future research lines include the integration with more sophisticated clustering techniques (different ensemble for different clusters), the exploitation of other meteorological data like humidity, temperature and wind speed, the correlation with predictions made in neighboring sites, working with different sampling time and prediction horizon to fit different applications.

Further, a better way to compute the seasonality can be investigated. One possibility can be using an empirical formula able to compute exactly the solar radiation that invests a certain zone in each period of the year. It would

improve every method studied in this project, because everyone is based on removing and re-adding the seasonality. So, a higher accuracy on the seasonality computed imply a higher accuracy of each predictions.

As stated in the introduction of the thesis, the accuracy of the prediction of solar radiation can affect drastically the efficiency of Photovoltaic Power generation systems. So, it can be evaluated the impact of the increased accuracy of the predictions of solar radiation by mean of the proposed methods on the prediction of PV power and so as a consequence the impact of the improvements of PV power predictions on the context of microgrids and energy market balancing.

Chapter 8

Bibliography

1. *Forecasting.* **Makridakis Sypros, Wheelwright Steven C, Hyndman Rob J.** Wiley Student Edition. 3rd ed., ISBN 978-0-471-532330.

2. **Yi, Lim Pei.** *Power management strategies for off-grid hybrid power systems.* s.l. : School of Electrical and Computer Engineering, Curtin University, 2011.

3. **Peder, Bacher.** *Models for efficient integration of solar energy [Ph.D thesis].* s.l. : Technical University of Denmark, 2012.

4. **Madsen Henrik, Nielsen Henrik Aalborg, Jonsson Tryggvi, Pinson Pierre,.** *Forecasting of wind and solar power production [Research.* s.l. : Technical University of Denmark, 2011.

5. *Predictive models for power management of a hybrid.* **Prema V, Uma Rao K.** Manipal, India : s.n., 2014. 2nd International Conference on Advances in Energy (ICAECT 2014).

6. *Stochastic modelling of solar irradiance on horizontal and vertical planes at a northerly location.* **Craggs C., Conway E. and Pearsall N.M.** 1999.

7. *Predicting solar radiation at high resolutions: A comparison of time series forecasts.* **G., Reikard.** 2009. Solar Energy.

8. *"Cloud motion tracking for short-term on-site cloud coverage prediction,".* **D. M. L. H. Dissawa, M. P. B. Ekanayake, G. M. R. I. Godaliyadda, J. B. Ekanayake and A. P. Agalgaonkar.** Colombo : s.n., 2017. Seventeenth International Conference on Advances in ICT for Emerging Regions (ICTer).

9. **Ming-Ching Chang, Yi Yao, Guan Li,Yan Tong, Peter Tu.** *Cloud tracking for solar irradiance prediciton.* Niskayuna, NY USA : GE Global Research Center, 2011.

10. *"Solar irradiation forecasting: state-of-the-art and proposition for future developments for small-scale insular grids".* **H. M. Diagne, P. Lauret, M. David.** 2012. WREF 2012- World Renewable Energy Forum.

11. *"Ensemble method based on ARIMA-FFNN for climate forecasting,".* **B. W. Otok, D. A. Lusia, Suhartono, R. Faulina, Sutikno and H. Kuswanto.** Langkawi : Department of Statistics Institut Teknologi Sepuluh Nopember , 2012. International Conference on Statistics in Science, Business and Engineering (ICSSBE).

12. *A Solar Time-based Analog Ensemble Method for Regional Solar Power Forecasting .* **X. Zhang, Y. Li, S. Lu, H. F. Hamann, B. Hodge and B. Lehman.** 2019, IEEE Transactions on Sustainable Energy, Vol. 10, p. 268-279.

13. *"An Improved Photovoltaic Power Forecasting Model With the Assistance of Aerosol Index Data".* **J. Liu, W. Fang, X. Zhang, and C. Yang.** 2015, IEEE Transactions on Sustainable Energy, Vol. 6, p. 434-442.

14. *"Solar Power Prediction Based on Satellite Images and Support Vector Machine".* **H. S. Jang, K. Y. Bae, H. S. Park, and D. K. Sung.** 2016, IEEE Transactions on Sustainable Energy, Vol. 7, p. 1255-1263.

15. *"A Hybrid Algorithm for Short-Term Solar Power Prediction-Sunshine State Case Study".* **A. Asrari, T. X. Wu, B. Ramos.** 2017, IEEE Transactions on Sustainable Energy, Vol. 8, pp. 582-591.

16. *Short-term solar power prediction using a support vector machine.* **Qiao, J. Zeng and W.** 2013, Renewable Energy, Vol. 52, p. 118-127.

17. *"Regional forecasts of photovoltaic power generation according to different data availability scenarios: a study of four methods,".* **J. G. d. S. Fonseca Junior, T. Oozeki, H. Ohtake, T. Takashima, and K. Ogimoto.** 2015, Progress in Photovoltaics: Research and Applications, Vol. 23, p. 1203-1218.

18. *"Regional forecasts and smoothing effect of photovoltaic power generation in Japan: An approach with principal component analysis," .* **J. G. da Silva Fonseca Junior, T. Oozeki, H. Ohtake, K.-i. Shimose, T. Takashima, and K. Ogimoto.** 2014, Renewable Energy, Vol. 68, p. 403-413.

19. **A.D., Watt.** *On the Nature and Distribution of Solar Radiation.* s.l. : U.S. Department of Energy, 1978.

20. **Williams, Dr. David R.** Wikipedia & Earth Fact Sheet. [Online] NASA Goddard Space Flight Center. [Riportato: 24 March 2019.] https://en.wikipedia.org/wiki/Sunrise_equation. https://nssdc.gsfc.nasa.gov/planetary/factsheet/earthfact.html.

21. **Beauducel, Francois.** MathWorks File Exchange - SUNRISE: sunrise and sunset times. [Online] 19 October 2017.

https://it.mathworks.com/matlabcentral/fileexchange/64692-sunrise-sunrise-and-sunset-times?s_tid=prof_contriblnk.

22. *A chronology of interpolation: from ancient astronomy to modern signal and image processing.* **Meijering, Erik.** 2002, Proceedings of the IEEE, p. 319–342.

23. *A Simple Proof of Rolle's Theorem for Finite Fields.* **Ballantine, C. e Roberts, J.** 2002, The American Mathematical Monthly, p. 72–74.

24. **Davis M.H.A., Vinter R.B.** Asymptotic analysis of prediction error identification methods. *Stochastic Modelling and Control. Monographs on Statistics and Applied Probability.* Dordrecht : Springer, 1985, p. 215-246.

25. **Box, G. E. P., G. M. Jenkins, and G. C. Reinsel.** *Time Series Analysis: Forecasting and Control. 3rd ed. Englewood Cliffs, NJ: Prentice Hall.* 1994.

26. —. *Time Series Analysis: Forecasting and Control. 3rd ed.* s.l. : Englewood Cliffs, NJ: Prentice Hall, 1994.

27. **Nau, Robert.** The mathematical structure of ARIMA models. [Online] Fuqua School of Business, Duke University, 13 December 2014. http://people.duke.edu/~rnau/Mathematical_structure_of_ARIMA_models--Robert_Nau.pdf.

28. **R. Deutsch: Estimation Theory. Prentice-Hall, Inc., Englewood Cliffs, N. J.**

29. **Zheng, Fengxia & Zhong, Shouming.** *Time series forecasting using an ensemble model incorporating ARIMA and ANN based on combined objectives.* 2011.

30. **Cortes, C. & Vapnik, V.** Machine Learning. *Support-vector networks.* s.l. : Kluwer Academic Publishers, 1995, p. 273–297.

31. *"Quantile-copula density forecast for wind power uncertainty modeling,".* **Bessa, R.** Trondheim, Norway, 2011, IEEE PowerTech, p. 1-8.

32. *"Statistical analysis of wind power forecast error,"* . **H. Bludszuweit, J. Dominguez-Navarro, and A. Llombart.** July, 2008, IEEE Trans. Power Syst, Vol. 23, p. 983-991.

33. *"Examining information entropy approaches as wind power forecasting performance metrics,".* **B.-M. Hodge, K. D. Orwig, and M. Milligan.** Istanbul,Turkey : s.n., 2012. IEEE 12th Int. Conf. Probab. Methods Appl. Power Syst.

34. *"On uncertainty of wind power predictions—Analysis of the forecast accuracy and statistical distribution of errors,"* . **Lange, M.** 2005. s—Analysis of the forecast accuracyandstatistical distribution oferrors," Trans. ASME-N-J. Solar Energy Eng., vol. 127, pp. 177–184,.

35. *"Recent Trends in Variable Generation Forecasting and Its Value to the Power System,"* . **K. D. Orwig, M. L. Ahlstrom, V. Banunarayanan.** July 2015. , IEEE Transactions on Sustainable Energy , Vol. 6, p. 924–933.

36. *Development of statistical time series models for solar power prediction.* **Rao, V. Prema and K. U.** 2015, ScienceDirect.