TÉCNICO LISBOA

# Solar Irradiance Forecast Using Artificial Intelligence Techniques

**José Diogo Marques do Rego**

Thesis to obtain the Master of Science Degree in

## Electrical and Computer Engineering

Supervisor(s): Prof. Rui Manuel Gameiro de Castro

### Examination Committee

Chairperson: Prof. Horácio Cláudio de Campos Neto
Supervisor: Prof. Rui Manuel Gameiro de Castro
Member of the Committee: Prof. Paulo José da Costa Branco

**May 2017**

To me and my Grandparents...

# Acknowledgments

I would like to leave a few words of thanks to those who have contributed to the accomplishment of this work.

First of all, I would like to express my gratitude to Professor Rui Castro, the scientific advisor of the present dissertation, for the teachings and the constant availability.

I would also like to thank Professor Luís Roberto for the suggestions and the way he followed the beginning of this work.

To my College Friends. For the good moments spent with you, for the support of everyone, and for making this one of the best decisions made in my life.

To Cláudia, Rita and Jorge,my companions from the Energy Area. Because they embarked with me in this master's degree, without your support and teamwork this barrier would have been much more difficult to overcome.

To Nobre, my first friend in the Capital. For all the words and counsels of a friend and for challenging myself to enter into other worlds that have made me the most complete Man that I am today.

To the Imaculados, Sofia, João, Gonçalo. For the support, affection, friendship, laughter and for making me feel always at home being 170 km from it. "To us, to the House and to our House".

To Flávio. For the truly wise advice worthy of a high school godfather.

To my Handball friends. For having increased a passion that was already inside me, for the moments of camaraderie, and for all the achievements conquered in your company. "AEIST".

To my SFA friends. For also helping me to grow as a person, to make me forget the bitterness of life, also to support me in everything and give me so many good stories. "SFA is our Great Love."

To Mariana. For all the support you gave me, for the motivation given in the most demanding hours, for your understanding, for your patience, for your smile, and for giving me reasons to be better every day.

To my Godmother. By sharing your experience, wisdom, and availability as long as I remember.

To my Sisters. For being the best gift that my parents could give me, for the motivation they always

give me and for encouraging me only with their existence.

And finally to my parents. For being the fantastic parents that they are, because they are my safe harbor, for being my best example, for all the support and strength they gave me, and for the patience they always had for me.

# Resumo

A importância das energias renováveis tem vindo a crescer a um ritmo acelerado, quer devido à necessidade de resolver problemas relacionados com as questões ambientais, quer como forma de ajudar a gestão, cada vez mais difícil, das redes eléctricas. As técnicas de Inteligência Artificial já mostraram a sua eficácia em tarefas de elevada complexidade (e.g. Regressão, Classificação, Previsão). Também no campo das Energias Renováveis estas ferramentas podem ser extremamente úteis, nomeadamente na previsão da Irradiância Solar. Neste trabalho foram desenvolvidos dois algoritmos em Matlab® de previsão de Irradiância Solar baseados em dois métodos de Inteligência Artificial, a saber: as Redes Neuronais Artificiais e o Método dos k-Vizinhos Mais Próximos. No processo de previsão, os modelos são treinados com subconjuntos do registo de um ano de Irradiância Solar na cidade de Lisboa para depois se fazer a previsão da próxima hora. Para se entender qual o melhor método para realizar previsões de Irradiância Solar, de entre aqueles que foram estudados, foi realizado um estudo comparativo entre os modelos, tendo em conta os erros de previsão e os tempos de simulação de ambos nas simulações feitas em diferentes situações.

**Palavras-chave:** ANN, KNN, energia solar, previsão, energias renováveis, aprendizagem automática.

## Abstract

The importance of renewable energies has been growing at a fast pace, both because of the need to solve problems related to environmental issues and as a way of helping the increasingly difficult management of electricity grids. The techniques of Artificial Intelligence have already shown their effectiveness in tasks of high complexity, namely, Regression, Classification, Forecasting. Also in the field of Renewable energies these tools can be extremely useful, in particular in the prediction of Solar Irradiance. In this work, we developed two algorithms in Matlab® of prediction of Solar Irradiance based on two methods of Artificial Intelligence, which are the Artificial Neural Networks and the K-Nearest Neighbors Method. In the forecasting process, the models are trained with subsets of the one-year Solar Irradiance register in the city of Lisbon and then the next hour's forecast is carried out. In order to understand the best method to perform predictions of solar irradiance among those studied, a comparative study between the models was carried out, taking into consideration the prediction errors and the simulation times of both models in the simulations made in different situations.



**Keywords:** ANN, KNN, solar power, forecasting, renewable energy, machine learning.

# Contents

# List of Tables

# List of Figures

# Acronyms List

| | |
|---|---|
| **ANN** | Artificial Neural Network |
| **CPV** | Concentrated Photovoltaics |
| **CSP** | Concentrated Solar Power |
| **ELA** | Ensembling Learning Algorithm |
| **EU** | European Union |
| **GHG** | Greenhouse Gas |
| **IEA** | International Energy Agency |
| **KNN** | K-Nearest Neighbors |
| **MAE** | Mean Absolute Error |
| **MAPE** | Mean Absolute Percentage Error |
| **Mtoe** | Million Tonnes of Oil-Equivalent |
| **RES** | Renewable Energy Sources |
| **RMSE** | Root Mean Square Error |
| **TFC** | Total Final Consumption |
| **TPES** | Total Primary Energy Supply |
| **UNFCCC** | United Nations Framework Convention on Climate Change |

# Chapter 1

# Introduction

Renewable Energy is becoming a technology and an increasingly viable alternative to the conventional Non-Renewable Sources of Energy. The constant threat of the Climate Changes thus requires the mankind to look for better and more efficient ways of producing electricity. Specially when all our basic needs are based on this type of energy.

The total amount of Solar Energy delivered by the Sun to the Earth is unimaginable. The San Francisco earthquake of 1906 reached a $7.8$ Richter's magnitude equal to $10^{17}J$ of released energy, which is the same amount of energy delivered by the Sun per second [1]. The total World Oil Reserves (about $1.7 \times 10^{22}J$) are exactly the same amount of energy received by the Earth in a day and a half. During an hour, the sun delivers to the Earth the same amount used in the human activities during a year ($4.6 \times 10^{20}J$).

After the big increase in the use of the Hydro and Wind Power, now it is the Solar Power that is showing some signs of increasing its use and it is one in which it has been invested most in order to improve its efficiency. The technology associated with the Solar Power is improving at a good pace as we can see by the appearance of new types of cell candidates to the replacement of the traditional and still expensive Crystalline Silicon Cells, like the Thin-Film Technologies, the Multijunction Cells or the Emerging Photovoltaic Technologies (i.e. Organic Cells, Dye-sensitized Cells or Perovskite Cells). These facts, coupled with the global need to reduce Greenhouse Gas emissions, make the Solar Power one of the most promising Renewable Energies.

The following sections will show in more detail the main reasons for choosing this theme, the main objectives of this work and the structure of this thesis report.

## 1.1 Motivation

Currently, Solar Power is a well-known Renewable Energy Source (RES) with a great potential. But his dependence on external meteorologic conditions turns this technology somewhat volatile and sometimes unattractive both at the energy and at the economic level.

Currently the use of Artificial Intelligence Tools is becoming more and more common due to their capacity of solving highly complex problems. The improvement in the computers and in the algorithms performance also helped at solving problems, not only in engineering, but also in many areas like medicine, finances and literature. That is why these tools are becoming more popular and can still make some progress and can be applied in more areas or problems. In Table 1.1 we present some examples of application of Artificial Intelligence Tools in different study fields.

Table 1.1: Fields of Application of Artificial Intelligence Tools

| Field | Description |
|---|---|
| Automotive | Automotive automatic guidance systems, automatic braking systems, misfire detection, virtual emission sensors |
| Banking | Credit application evaluators, cash forecasting, firm classification, exchange rate forecasting |
| Electronics | Code sequence prediction, process control, chip failure analysis, nonlinear modeling |
| Medical | Breast cancer cell analysis, EEG and ECG analysis, optimization of transplant times, emergency room test advisement |
| Speech | Speech recognition, speech compression, vowel classification, text to speech synthesis |
| Telecommunications | Image and data compression, automated information services, real-time translation of spoken language, customer payment processing systems |

In the case of RES these artificial intelligence tools are used to perform forecasts of the energy produced in a Power Station or even forecasts of the behavior of weather conditions. These forecasts have a big importance in the Energy Market. The problem of variability and unpredictability of the Solar Irradiance that reaches the Earth surface is well-known. Thus, an accurate forecast of this variable can promote a better planning and operation of Power Distribution at economic level or energy production level, either by making alternative arrangements for conventional power and overall schedules or committing the right amount of energy resources and reserves in order to reduce the power system operational costs. In addition to these advantages a good forecast can help to reduce the impact of PV output uncertainty

on the grid, improve the system reliability, maintain the Power Quality and increase the penetration level of PV Systems. The forecasts generated by these tools can also be useful for studying the viability of solar production projects in a given location.

There were several relevant reasons for choosing this theme of the thesis. The evaluation of the importance that this kind of technology has but also the potential that it can bring to the world necessities, the understanding of Portugal's current situation on the Energy Sector, the principal characteristics of the resource used in this kind of technology and the main advantages of forecasting Solar Irradiance became key factors in the theme framework. Some of these topics will be better explored in Chapter 2.

## 1.2 Thesis Objectives

Due to the importance of a good Solar Forecast referenced before, the main objective of this thesis is to understand what method, among the studied methods, presents the best performance in Solar Forecasting.

To do that, first we will investigate on the available Artificial Intelligence techniques and their applications on Solar Irradiance Forecast. Next, we will implement a Solar Irradiance Forecast model using Matlab® Language and the techniques studied before. Lastly, in order to validate the accuracy of the developed models they will be submitted to experimental tests applied to a real case study and the results of the tests will be compared.

In order to make the best comparison possible both methods need to be well adapted to the problem proposed and as such, to overcome some weaknesses and problems characteristic of each model, before the comparison of the performance of both models, it is necessary to perform a slight optimization of both methods.

## 1.3 Thesis Structure

This work gives a theoretical and a practical vision of the study made on Solar Forecasting and is divided in several chapters. The next chapter presents a current state of play of the Solar Energy around the world and in our country (e.g. policies taken around the world, statistics of energy produced) as well as some characteristics and behaviors of this RES. In this chapter we can realize the growth of the importance of this kind of technology and understand why we should invest in it (Chapter 2).

Before the comparison, in Chapter 3 we will present a theoretical point of view of the two Models as well as the model used as reference, the Persistence Model. The main characteristics, their capacity and algorithms of learning or forecast information and the principal parameters of each model are some of the topics to be presented in this chapter.

The Chapter 4 presents all the information about the Case Study. Firstly, all the details about the Data used in the tests will be explained, then how the chosen models will be applied in this case study (how the tests were made, how the information data were stored to train the models, etc). In the end of the Chapter the mathematical tools used to evaluate the performance of the models in each test will be provided.

The next chapter (Chapter 5 - Results) will present the outcomes of all the tests and simulations made during the study. It starts with the results of the tests made to determine the best parameters on each Model and after that we will present the results for the comparative study between the two chosen models of this work.

Finally, the last chapter (Chapter 6 - Conclusion) presents the conclusions of this work, the comparisons of the performance of each model after getting the best parameters and some suggestions for further work on this area.

# Chapter 2

# Framework and State of The Art

This chapter explores the importance of Solar Power to the world, contextualizing its use and the policies taken by governments, as well as the typical characteristics of this type of RES. The first section (Section 2.1) talks about how the use of Solar Power benefits and helps in solving the Climate Change problems and what has been done by the international entities to increase the use of Solar Power and other RES. The Section 2.2 makes a similar analysis but in this case concerning the Portuguese case. The last section (Section 2.3) explores the existing work in the area of forecasting models.

## 2.1 The Importance of the Solar Power and Global Overview

Since the Industrial Revolution that the use of Fossil Fuels was of big importance for the Energy Supply in every sectors (e.g. Industry, Transport, Domestic use). After several studies about the effects of the use of Fossil Fuels, Humanity became aware of the harmful effects on the environment and on living things caused by burning this kind of fuels to produce energy. The energy production process using Fossil Fuels releases large quantities of Carbon Dioxide ($CO_2$ is the most anthropogenic GHG) into the atmosphere, which increases the Global Earth Temperature (Figure 2.1) [2]. As a consequence, it causes the melting of polar ice caps, and increases the average level of the sea water (Figure 2.2) [3].
With this evidence of the danger of overuse of Fossil Fuels, Governments and Societies have taken a more environmentalist position by adopting measures and developing new technologies to reduce the dependence on Fossil Fuels and the damage to the Environment. A good example of this is the Kyoto Protocol [4]. The first steps towards the creation of the Kyoto Protocol were given in 1992 at the United Nations Framework Convention on Climate Change (UNFCCC) but the Protocol was only adopted in 1997 by the UNFCCC parties and entered into force just in February 2005. The main objective of this treaty is the reduction of the emission of the anthropogenic GHGs, the biggest cause for the Global Warming, through the cooperation between the parties. This is supposed to be achieved through the establishment of binding commitments to reduce the emissions, prepare policies and measures to reduce these emissions, promote the use of RES and other clean technologies, reform the Energy and Transport Sectors both in terms of reducing emissions and in terms of energy efficiency and protect

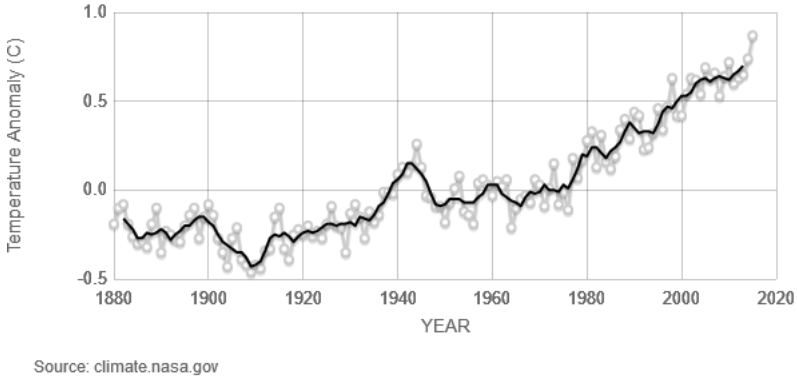



Figure 2.1: Evolution of World's Annual Average Temperature Anomaly from 1880 to 2015. The points represent the Annual Average and the black line represents the 5 year Mean. Source: NASA's Goddard Institute for Space Studies, GISS, [2].

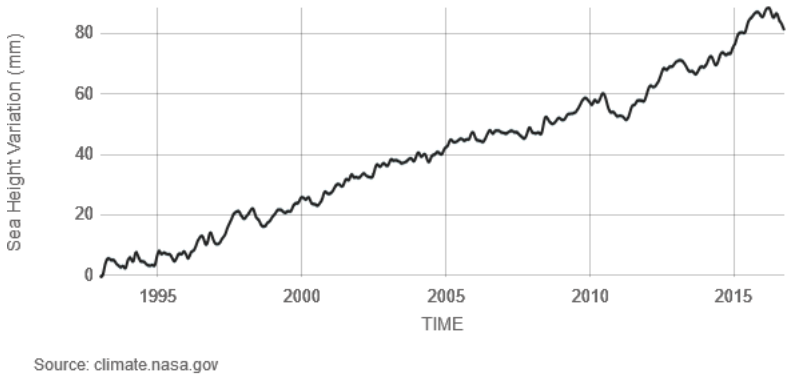



Figure 2.2: Evolution of World's Average Sea Height Variation from 1993 to September 2016. Source: NASA Goddard Space Flight Center, [3].

forests and other carbon sinks.

In order to meet the expectations set by the Kyoto Treaty many plans, treaties, legislations and measures have been taken over the years by the International Entities, for example, the EU 2020 Package or, more recently, the Paris Agreement. The first one is, simultaneously, a part of the principal objectives of Europe 2020 strategy for smart, sustainable and inclusive growth and a set of legislation to ensure that the EU members meet the climate and energy targets by the year of 2020 [5]. The three main targets of this Package are: 20% cut in GHG's emissions to 1990 levels, 20% of EU's energy from RES and 20% improvement in Energy Efficiency. The Paris Agreement emerged at the UNFCCC held in Paris (November-December 2015) after the negotiations on more measures to be taken after the year of 2020 and it was the first universal and legally binding global climate deal [6]. The principal objective of this agreement is to limit the global warming to well below $2°C$ through several actions in different areas like Mitigation of Emissions (i.e. undertaking rapid reductions in accordance with the best available science), Transparency and Global Stocktake (i.e. reporting between the members and the public the progress and results on reaching their targets), Adaptation (i.e. strengthening societies ability to deal with the impacts of climate change), Loss and Damage (i.e. averting, minimizing and addressing loss and damage associated with the effects of climate change), Role of Cities, Regions and Local Authorities

(i.e. upholding and promoting regional and international cooperation) and Support (i.e. countries are encouraged to provide or continue to provide support among them voluntarily).

Due to this global awareness of the Global Warming effects, new technologies have emerged in the international scenario and the Solar Power is a good example of this. Started by being a niche market technology with small applications but it quickly became a well-known source of electricity for the entire world. Due to advances in the technology present in photovoltaic panels and cells, and the advance of the production techniques, the cost of the PV cells has also fallen sharply. For example, the price of Crystaline Silicon Photovoltaic Cells has changed from around 70$\$/W$ in 1977 to around 1$\$/W$ in 2010, reaching nowadays, in some cases, values under the dolar per Watt of capacity [7]. This evolution, coupled with the fact that the sun is a free energy source and that more and more governments are adopting feed-in policies to encourage society to use RES, make this source of energy very relevant and an very attractive alternative to the conventional forms of energy.

| | | 2014 | 2015 |
|---|---|---|---|
| **INVESTMENT** | | | |
| New investment (annual) in renewable power and fuels | billion USD | 273 | **285.9** |
| **POWER** | | | |
| Renewable power capacity (total, not including hydro) | GW | 665 | **785** |
| Renewable power capacity (total, including hydro) | GW | 1,701 | **1,849** |
| Hydropower capacity | GW | 1,036 | **1,064** |
| Bio-power capacity | GW | 101 | **106** |
| Bio-power generation (annual) | TWh | 429 | **464** |
| Geothermal power capacity | GW | 12.9 | **13.2** |
| Solar PV capacity | GW | 177 | **227** |
| Concentrating solar thermal power | GW | 4.3 | **4.8** |
| Wind power capacity | GW | 370 | **433** |
| **HEAT** | | | |
| Solar hot water capacity | $GW_{th}$ | 409 | **435** |
| **TRANSPORT** | | | |
| Ethanol production (annual) | billion litres | 94.5 | **98.3** |
| Biodiesel production (annual) | billion litres | 30.4 | **30.1** |

REN21 *Renewables 2016 Global Status Report*

Figure 2.3: Indicators for Investment in RES and World's Installed Capacity in 2014 and 2015. Source: REN21, [8].

According to the most recent REN21's report [8] 2015 was an outstanding year for Renewable Energies, especially for Solar Power and Wind Power. The sharp drop in Fossil Fuel prices was remarkable, the reduction of prices of Renewable Power long-term contracts to the lowest levels ever seen, the increasing of the importance of energy storage, the historic Paris Agreement and the fact that the net investment in power capacity additions from RES surpassed the investment made in additions using Fossil Fuels for the $6^{th}$ consecutive year. The fast growth of the use of RE can be explained in this report by several reasons such as: the improving of competitiveness of the renewable technologies costs, the creation of dedicated policies by governments and more funding for such projects (Figure2.3), the increasing

of the concern for environmental and for energetic security matters, and the great need for energy in developing and emerging economies as well as access to modern kinds of energy. This fast growth and these reasons created conditions to the appearance of new markets all over the globe for projects in both centralized and distributed renewable energy. In addition to the remarkable achievements in the field of markets and investment growth, the report also highlights the improvements made with regard to the existing technologies, such as, the advances made in the RE Technologies and in the Energetic Efficiency, the increase in the use on Smartgrid Technologies, the progress made in hardware and software to support the integration of RE and the progress made in Energy Storage.

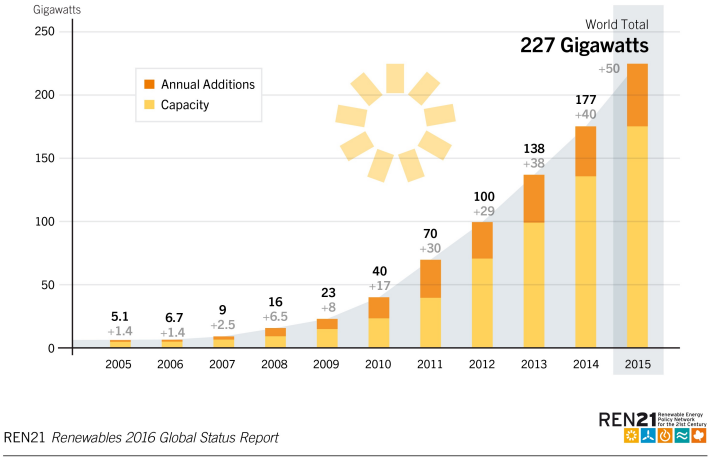



Figure 2.4: Evolution and annual addiction of Solar PV Installed Capacity between 2005 and 2015. Source: REN21, [8].

Statiscally we can observe the progress made by the RE and the Solar Energy in particular. According to the REN21's report in 2015 RES reached a 23.7% share of Global Electricity Production and the total addition of Renewable Power Capacity is equal to 8.7%, where the contribution given by Solar PV is equivalent to 1.2% and 2.9% respectivily. This kind of RES is simultaneously the technology with the biggest power capacity increase from 2014 to 2015 and the technology with biggest growth rate from 2010 to 2015. Comparing the general performance of the International Energy Agency (IEA) Members we can observe that the biggest users of RES are the Norwegians due to their high percentage of Hidro Power (Figures 2.5 and 2.6).

Basing our analysis only on the Solar Power we can observe that in a period of 10 years, from 2005 to 2015, the global power capacity evolved from 5.1 to 227 GW of Solar PV (Figure 2.4) and from around 100 to 435 GW of Solar Thermal Power (Figure B.8). Currently, China is the leading country in the use and investment in Solar PV closely followed by Japan (Figure B.9 and B.10). Other countries have a big influence in this RES like the United States for being one of the major investors in Solar PV in 2015 or Germany for being the second country with more installed capacity by the end of the same year. Finally, analyzing the amount of direct and indirect jobs in Renewable Energy worldwide we can observe that the Solar Energy is the technology that creates a greater number of jobs (excluding the jobs in large scale hydro power) which is a great advantage on economic level (Figure 2.7).

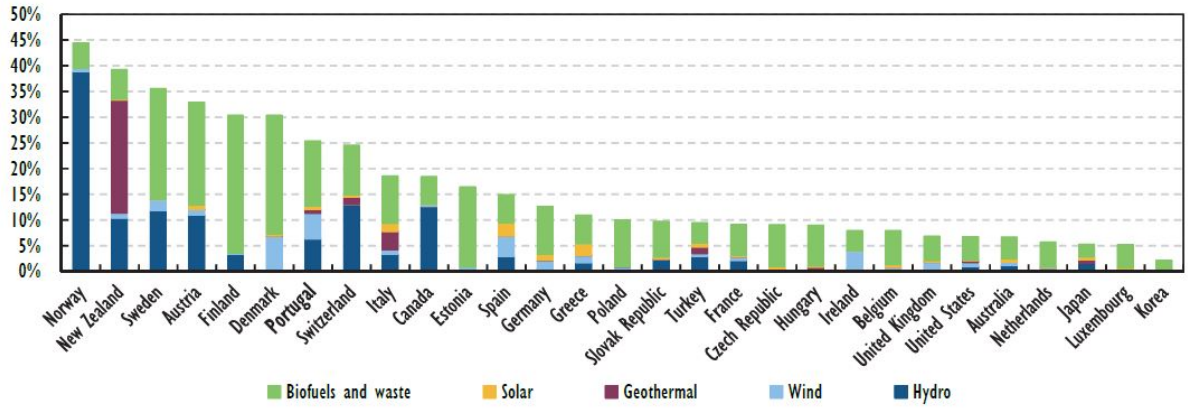



Figure 2.5: Ranking of all IEA Member countries in order of renewable energy produced as a percentage of Total Primary Energy Supply (TPES) in the year of 2014. Note: Data are estimated. Source: IEA (2015), Energy Balances of OECD Countries 2015, www.iea.org/statistics/,[9].

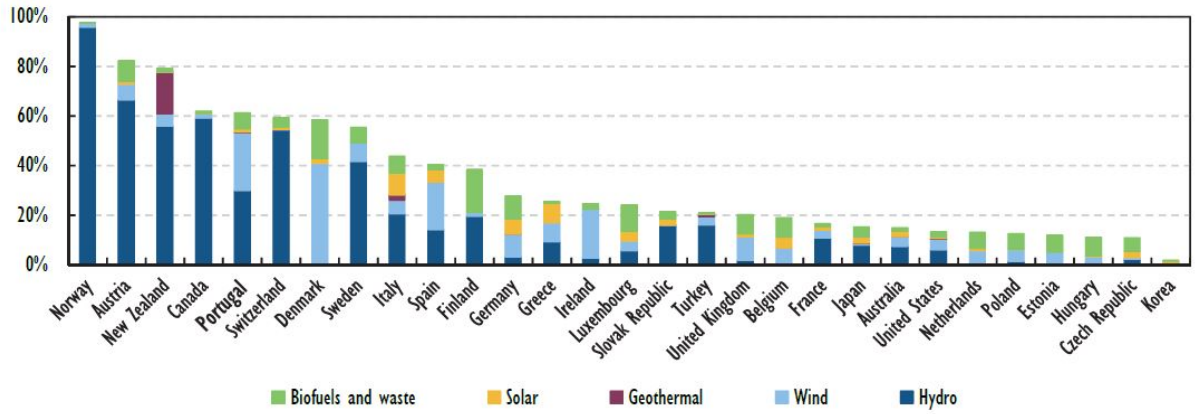



Figure 2.6: Ranking of all IEA Member countries in order of renewable energy produced as a percentage of all energy produced in the year of 2014. Note: Data are estimated. Source: IEA (2015), Energy Balances of OECD Countries 2015, www.iea.org/statistics/,[9].

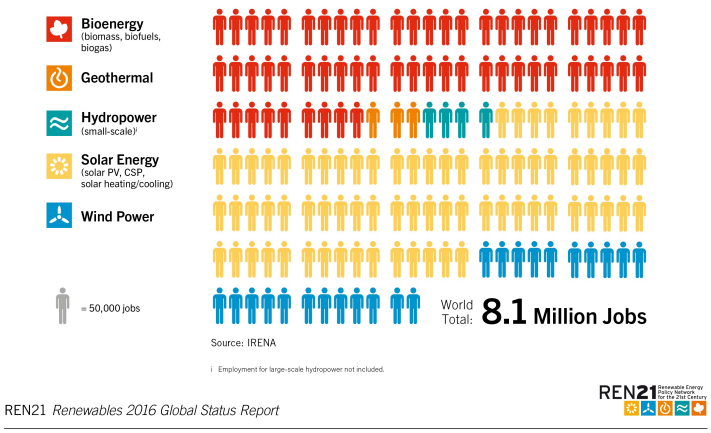



Figure 2.7: Distribution of jobs in renewable energy market by type of technology. Source: REN21, [8].

In Annex B can be consulted more comparative statistics of Renewables Market.

## 2.2 Portugal's Overview

The Portuguese case is not different from the other countries around the world and to reach the objectives set by the recent Paris Agreement changes to the current energy policy and more investment in Renewable Energies must be made by the Portuguese Government. Nevertheless, the last International Energy Agency (IEA) report about the Portuguese performance in the energy sector is very positive.
According to [9], in 2014, Portugal produced a total of 5,6 million tonnes of oil-equivalent (Mtoe) of energy. Since 2004 the production has increased 44,4% and the average production in this 10 year period is equal to 5,0 Mtoe. As we can see in Figure2.8 the production over the years is highly volatile. This is due to the fact that the production is entirely from Renewable Sources. The biggest share of the total production of 2014 correspond to Biofuels and Waste energy production (52,2%) while the Solar share is about 2,3% (more 20% than 2013).

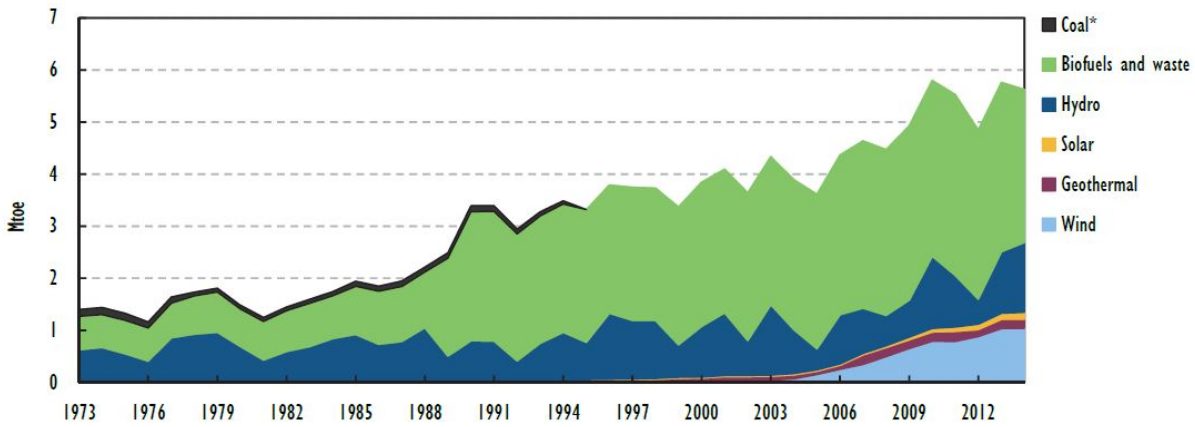



Figure 2.8: Energy Production by source in Portugal between 1973 and 2014. Source: IEA (2015), Energy Balances of OECD Countries 2015, www.iea.org/statistics/,[9].

Another indicator of a positive performance of the Portuguese performance in the energy sector is the comparison between the Total Primary Energy Sector (TPES) and the Total Final Consumption (TFC). TPES is the total energy supplied that is consumed domestically. The TFC is the total energy consumed by end-users (e.g. in the form of electricity, heat, gas, oil products), excluding fuels used in electricity and heat generation and other energy industries such refining. The records of these values from 1973 until 2014 are represented in Figure 2.9 and 2.10.
According to the report, Portugal reached a TPES of 21,1 Mtoe in 2014 (less 18,3% than the TPES of 2003) and a TFC of 16,2 Mtoe in 2013 (less 18,9% than the TFC of 2003). The RES contribute to 25,4% of total TPES. Solar Power is included in the portion of Renewable Energies and correspond to 0,6% of total TPES. This share, although at a slower pace compared to other technologies, had a great expansion since 2004 increasing from 0,1% of TPES to 0,6% of 2014 (Figure 2.11).
Comparing with the other country members of the IEA, Portugal is in $7^{th}$ place in the list of the countries with the biggest share of renewables in TPES, with the second highest share of wind behind Denmark, the fifth highest in geothermal and the seventh highest in Hydro and Solar (Figure 2.5).
With regard to the renewable energies share in electricity generation Portugal is well ranked among the

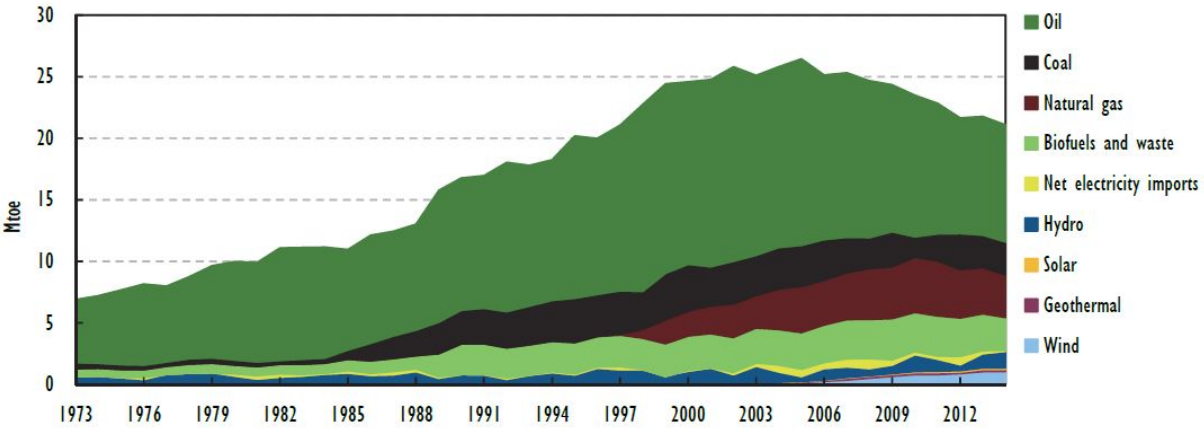



Figure 2.9: TPES by source in Portugal between 1973 and 2014. Source: IEA (2015), Energy Balances of OECD Countries 2015, www.iea.org/statistics/,[9].

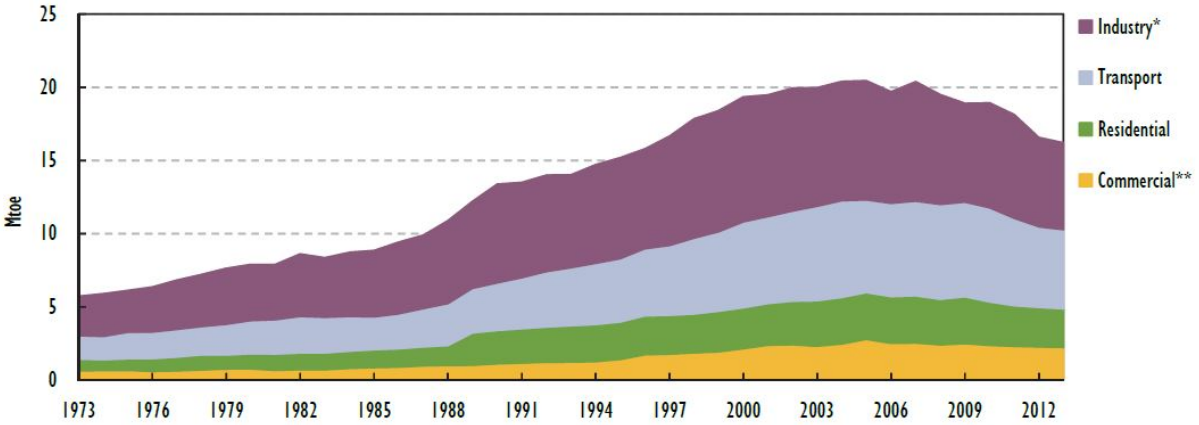



Figure 2.10: TFC by sector in Portugal between 1973 and 2013. *Industry includes non-energy use. **Commercial includes commercial and public services, agriculture, fishing and forestry. Source: IEA (2015), Energy Balances of OECD Countries 2015, www.iea.org/statistics/[9].

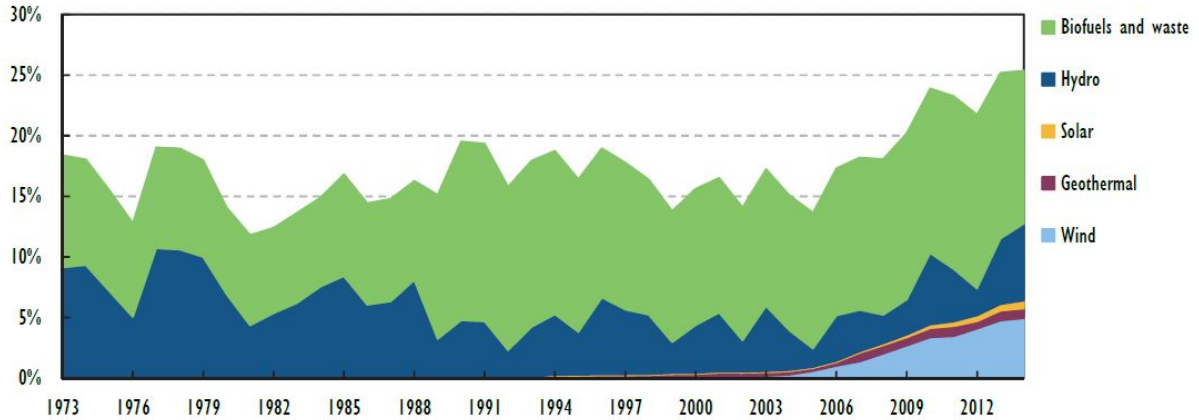



Figure 2.11: Energy produced in Portugal between 1973 and 2013 by renewable sources as a percentage of TPES. Note: Data are estimated for 2014. Source: IEA (2015), Energy Balances of OECD Countries 2015, www.iea.org/statistics/,[9].

rest of the members of the IEA (Figure 2.6). Portugal is ranked the fifth-highest among IEA members in the total amount of the annual production. 61,3% of total generation (31,9 TWh in 2014) is from renew-

able origin. This share has increased 28,1% since 2004. Hydro, Wind Power and Geothermal are used only in electricity generation and correspond of 30%, 23,3%, and 0,4% of total generation respectively, 27% of Biofuels and Waste go into electricity and heat production and correspond to 6,4% of the total generation, and Solar Power is consumed not only in Power Generation but also in Households and Businesses corresponding to 1,2% of total electricity Generation.

The Portuguese share of Wind is, like the share in TPES, the second highest among the IEA members, just behind Denmark, is the fifth-highest in the Geothermal share, the seventh-highest in the Hydro share and the fourteenth-highest in the Solar Power among IEA members.

Like the other members of the European Union (EU) Portugal established objectives for the Energy sector to contribute to the main objectives of the EU for this Sector (20% of reduction in greenhouse gas emissions close to the levels of 1990, 20% share of renewable energy produced on total final consumption and 20% of reduction of the primary energy consumption relatively to the projections for 2020). To achieve these objectives the EU established guide lines that require all the state members to present periodic plans of action for Energetic Efficiency and for Renewable Energies. Since 2008 Portugal has been developing different plans to improve Energetic Efficiency (National Energy Efficiency Action Plan, NEEAP) and the use of RES (National Renewable Energy Action Plan, NREAP).

The first NREAP submitted in 2010 by the Portuguese government to the European Comission, presented several different measures to promote RES (e.g. a pilot zone for wave technologies, solar energy technology demonstration projects and several photovoltaic power stations in the south of the country) [10]. Due to the significant changes in the macroeconomic situation in Portugal and Europe and requested review of the NEEAP by the EU, the Portuguese government was forced to review his NEEAP and NREAP programs. In terms of results, in 2010 Portugal had already reached good records of good records of RES incorporation in gross energy consumption. About 34,5% of the Heating and Cooling Sector was from RES (target of 30,6%), 41,1% of the Electric Sector was from RES (target of 55,3%), in the transport sector 5,5% of gross energy consumption of this sector was from RES (target of 10%) and in the total gross energy consumption of the country 24,6% was from RES(target of 31%). Besides that, the Portuguese Government was forced to elaborate a new NREAP (NREAP 2020) due to the significant changes in the National and European macroeconomic environment (mainly the declining of energy consumption and funding constraints) and the need imposed by the EU Energy Efficiency Directive to review the current NREEP.

As in the 2010's NREAP, the Portuguese main objective is to achieve 31% share of RES energy (the fifth-highest in EU) in the gross energy consumption and 10% in transport energy consumption. Relatively to the 2010's NREAP, the prediction of installed capacity for 2020 reduced 18% but thanks to the effect of the reviewed NREEP the expectations for 2020 for the share of electricity from RES and for the final share in gross energy consumption had increased relatively the 2010's NREAP expectations (about 60% against the previous 55% and about 35% against the previous 31% respectively). The main measures already present in 2010's NREAP but which were reviewed are:

- redefinition of the support mechanisms associated with emerging technologies;

- revaluation of the goals associated to the Concentrated Solar Power (CSP) and Concentrated Photovoltaics (CPV) Power Stations, due to still high cost on energy production;
- review of the goals and the objectives of micro and mini-generation;
- changing the high investment policies in the H&C sector for regulatory policies;
- maintain the effort of promotion of the use of biofuels and other renewable fuels in the Transport Sector;
- stimulate the development of energetic use from biomass (mostly forest origin), in particular on the support to the biomass equipments for heating and for Hot Water and Sanitation in the domestic and in the public service sectors.

In the Electricity Sector, Portugal presented a total RES installed capacity of 10.623 MW in 2011 which represents an increase of 110% since 2000 and an increase of 10% relatively 2010. This installed capacity allowed to generate 48% of total gross domestic electricity production while in 2000 this share corresponded to only 31%. For the period between 2010-2020 the latest NREAP anticipated an average annual growth of 5% of RES installed capacity and 1% of RES electrical generation. To fulfill the objectives for 2020 in the electric sector it will be necessary, according to the most recent NREAP, an increase of 59,6% of RES installed capacity in comparison to 2010 and an increase of 29% of electric energy generated relatively to the values of 2011. The NREAP policies for Solar Power consist of:

- the install of about 250 MW of Mini-Generation distributed in the services sector (Schools, Public Infrastructures and Big Distributors) the install of about 80 MW of Micro-Generation and the construction of about 50 MW od Solar Thermoelectric until 2020;
- the construction of Photovoltaic Power Stations with bigger power values it will depend on the evolution of the price of the technologies, but in 2010 70 lots were awarded for the construction of Photovoltaic Power Stations in the value of 140 MW of installed capacity;
- the construction of experimental CSP and CPV units to demonstrate the viability of technology.

In Figures 2.12 to 2.15 are visible the comparisons between the expectations of NREAP and what happened in Portugal since the implantation of NREAP and NEEAP.

These reasons and statistics prove that Solar Power is an important RES and has a great margin of evolution in these times. For Portugal, we may consider that we are in a good position regarding the use of RES and undeniable what has already been achieved until now, but it still has a long way to go.

## 2.3 State-of-Art on Solar Irradiance Forecasting

This section presents an overview of the most used approaches in Solar Irradiance Forecasting, as well as, the state-of-art of the models studied and compared in this work.

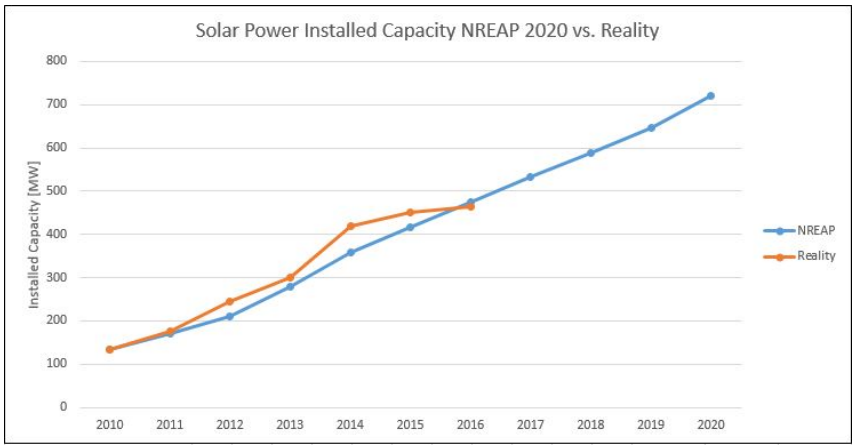



Figure 2.12: Comparison between NREAP and Reality Solar Power Installed Capacity in the period 2010-2020. Source: Diário da República, Nº70, 10 de Abril de 2013, [10][11].

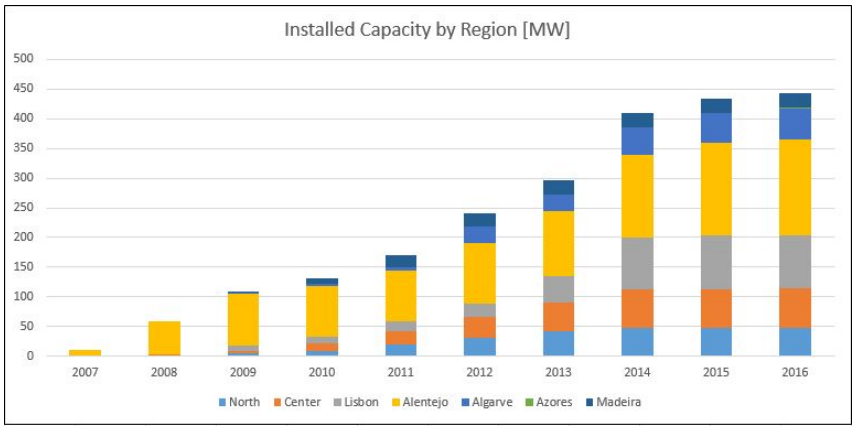



Figure 2.13: Solar Power Installed Capacity in Portugal by Region in the period of 2007-2016. Source: APREN records, [11].

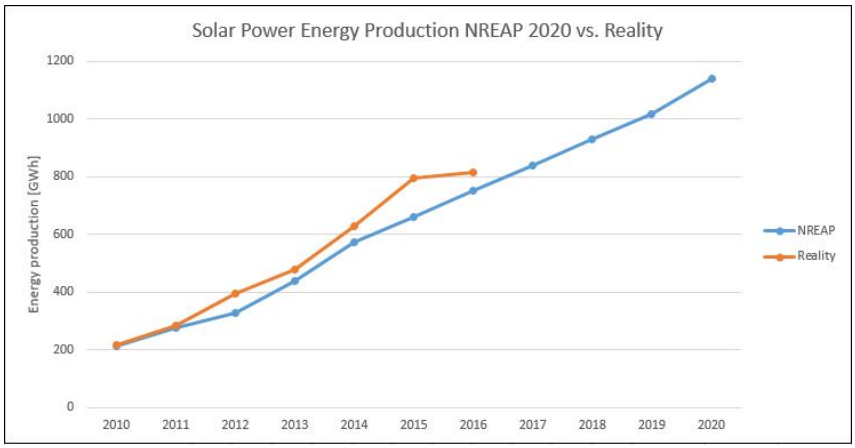



Figure 2.14: Comparison between NREAP and Reality Solar Power Energy Production in the period 2010-2020. Source: Diário da República, Nº70, 10 de Abril de 2013, [10][11].

Several methods of Solar Power and Solar Irradiance Forecasting have been developed over the years. In [12] the authors presented a solution combining a Support Vector Machine (SVM) System and a

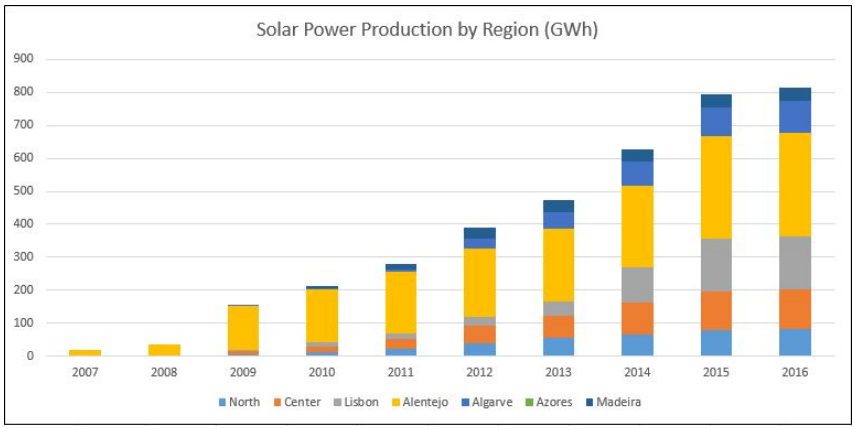



Figure 2.15: Solar Power Production in Portugal by Region in the period of 2007-2016. Source: APREN records, [11].

process of data series classification by meteorological types. The first step is to classify the PV power output data in 4 different groups (Cloudy, Foggy, Rainy and Sunny) and then the 4 SVM models will be established based on previous 4 groups. It has also been tried a more stochastic model as we can see in [13]. Historical Similar Mining (HISIMI) is a method that performs a search of the historical database of the model in order to find and use the most similar database cases with the current case. During this search, it is assign different weight values according to the degree of similarity between the historical and the current case. After this search, with the help of probability values and probabilistic functions, the model can compute the values of the predictions of forecast uncertainty and consequently the point forecast for each future instant. In [14] the authors suggested another combination of two different methods, this time the use of Wavelet Transform along with an Adaptive Neuro-Fuzzy Inference System. The wavelet technology is used to solve the problem caused by the data discontinuities and non-periodicity in the change of them. The proposed algorithm starts with the decomposition of the original data into number of wavelets coefficient signals. Then, these coefficients are used as input vectors of the ANFIS method that returns new coefficients. These coefficients recombine again using the same wavelet transform in order to predict the future solar irradiance.

The first works for the field of Neural Networks appeared in the late 19$^{th}$ and early 20$^{th}$ centuries as a result of an interdisciplinary work of physicists, psychologists and neurophysiologists. These works only includes general theories of learning, vision or conditioning, and does not make any specific mathematical models of neuron operations. These theories gained strength when in 1940's Warren McCulloch and Walter Pitts showed a modern view of Neural Networks [15]. In their work is showed that networks of artificial neurons could compute any arithmetic or logical function, which for many is acknowledge as the start of the neural network field. Despite this, the first application of an artificial neural network arose only at the hands of Frank Rosenblatt in the late 1950's with the invention of the Perceptron Network and Associated Learning Rule [16].

Since then, a lot of work have been done and this method started to be applied in the Solar Forecast-

ing. In [17] the authors compared the performance of three different ANN Architectures. The objective of their work was to find the best architecture to make a 24 hour forecast using data from the records of the previous days from the day to predict and from records of the same days and the next ones to those of the previous year. They compared the FFNN architecture with network with different transfer function (Radial Basis Function Neural Network, RBNN) or with a different network design (Recurrent Neural Network, RNN). They concluded that the performance of both architectures are similar although the RNN and RBNN slightly outperformed the FFNN in a few months.

In [18] the authors proposed and analyzed the performance of an ANN architecture (Multi-Layer Perceptron) in forecasting daily irradiation prediction. This analysis was made by comparing this method efficiency with the performance of other known algorithm (e.g. Naïve Predictor, Markov Chains, Bayesian inferences, KNN and ARIMA). First they tested the methods with non-stationary time series and they found that the ANN and the AR algorithms presented better predictions. Then they verified if a pre-processed data treatment would allow a better approximation between the predicted data and the measured data. Testing only with ANN, AR and ARMA, the results showed us that this pre-processed treatment would reduce the forecast error of about 5%-6% comparing with classic methods.

In [19] the authors proposed a solution that combines the well-known ANN method with a Wavelet Transform (WT). The application of WT helps to minimize the unpredictable changes in Solar Irradiance by filtering the time-series data. This proposed approach improved significantly the accuracy, efficiency and performance of the existing ANN Architectures. As we can see by the results disclosed in this work WT+RBFNN presented lower forecast for all seasons comparing with all other tested alternatives.

The Nearest Neighbors Method were first introduced theoretically by Evelyn Fix and J.L.Hodges, Jr in [20]. It took about 15 years, with the evolution of computer technology, so that the ideas of Fix and Hodges became feasible. The first generalized KNN decision rule only was explored in 1970 in the work of E.Patrick and F.Fischer[21].

In [22] the authors showed that, despite being more used in classification problems, the KNN method can be used to solve regression problems. In this case, they applied the KNN to solve a rainfall and runoff analysis of a known watershed. During their work, they compared the performance of three different methods (the Sacramento Method, ARMAX and the KNN). Analyzing the sum of residual square errors over the prediction of the three methods they find that the Sacramento Method were clearly the worst of the three models studied, and the KNN performance were quite similar to the ARMAX performance for this case study.

In [23] is proposed a variation on KNN combining a Genetic Algorithm (GA) to estimate the weights of the model. The performance is compared with the ANN. Tested in a 24-hour energy price forecasting problem. The authors concluded that the combination of the KNN with the GA is better than ANN. Lower

forecasting errors for the KNN+GA solution.

In [24] they applied the KNN method to make wind power forecast. Just like [23], the authors combined two different tools to obtain a forecast. In this case, they used the KNN algorithm to obtain the k closest historical examples with characteristics similar to the future weather conditions. After finding the k closest cases they use a Kernel Density Estimation method to develop a wind power predictive density. In this work, they proved that the proposed method works well for wind power forecasting either providing point or probabilistic forecasts.

In [25] the authors developed a different forecasting methodology for KNN method. In this work, they forecast GHI and DHI using local irradiance records and sky images. They capture pictures of the sky pointing the camera to zenith, filter the unnecessary information of the picture (e.g. trees, buildings) and convert the 8-bit RGB channels of the processed images into floating points that are going to be used in the forecasting process. Results reveals significant forecast improvements achieved by the KNN over a reference persistence forecast, but, contrary to the expectations, the inclusion of sky imagens into the KNN model does not improve the performance in a significant way. The explanation for this event lies in the fact that this inclusion has more impact on days with clouds in the sky.

# Chapter 3

# Forecasting Models

This chapter will explore the models studied in this work mathematically and theoretically, starting with the Persistence Model, the reference model used to evaluate the performance of the forecasts, followed by the definition of the models studied and compared in this work, the Artificial Neural Networks Model and the K Nearest Neighbors Model.

## 3.1 Persistence Model

The Persistence Model is the simplest forecasting model and is usually used as baseline or comparison for other models. During the study of new forecasting models, the performance of these models will be acceptable if they perform better than the Persistence Model. This Model calculates future time series values with the premise that all influential condition does not change between time $t$ and time $t + \Delta t$. For this work, this premise is applied to the evolution of irradiance and the model considers that the irradiance for time $t + \Delta t$, or the next irradiance value, is equal to irradiance for time $t$, or the actual irradiance value (example in Figure 3.1). Mathematically this can be represented by the equation:

$$I(t + \Delta t) = I(t) \tag{3.1}$$

where $I(t + \Delta t)$ is the irradiance for future time (where $\Delta t$ can be any time interval) and $I(t)$ the current value of irradiance.

## 3.2 Artificial Neural Networks (ANN)

The Artificial Neural Networks (ANN) Model is an Artificial Intelligence method based on the human capacity of learning and adapt his way of thinking through his obtained life time experience [26]. This method is capable of compute nonlinear modeling without knowing in the beginning the relation between input and output variables thanks to his nonlinear data-driven structures.

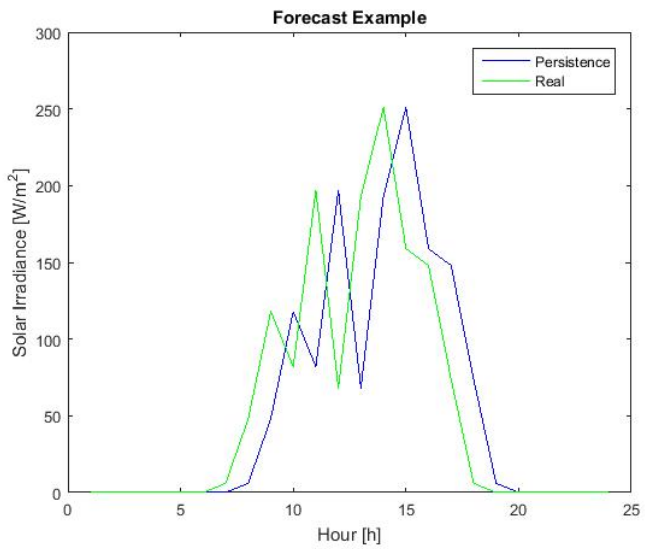
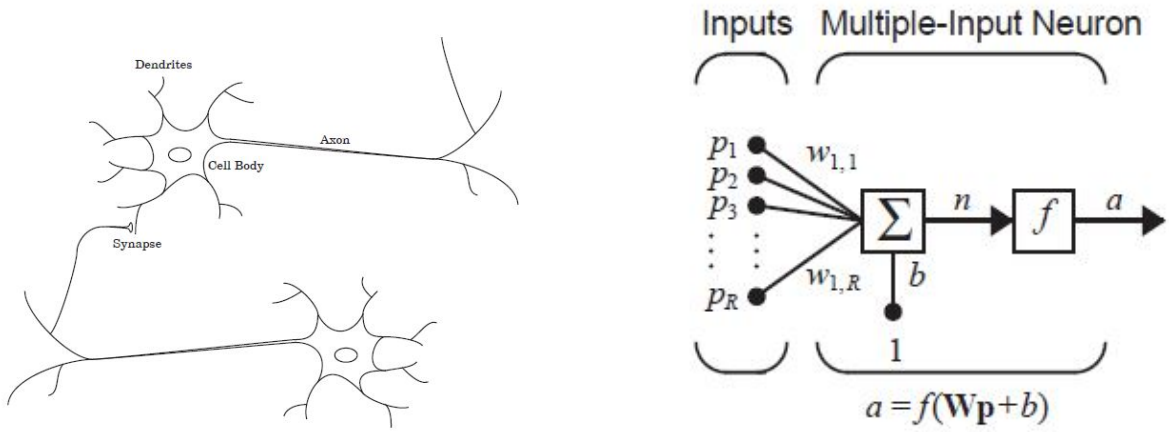



Figure 3.1: Solar Forecast example to explain the Persistence Model Forecast.



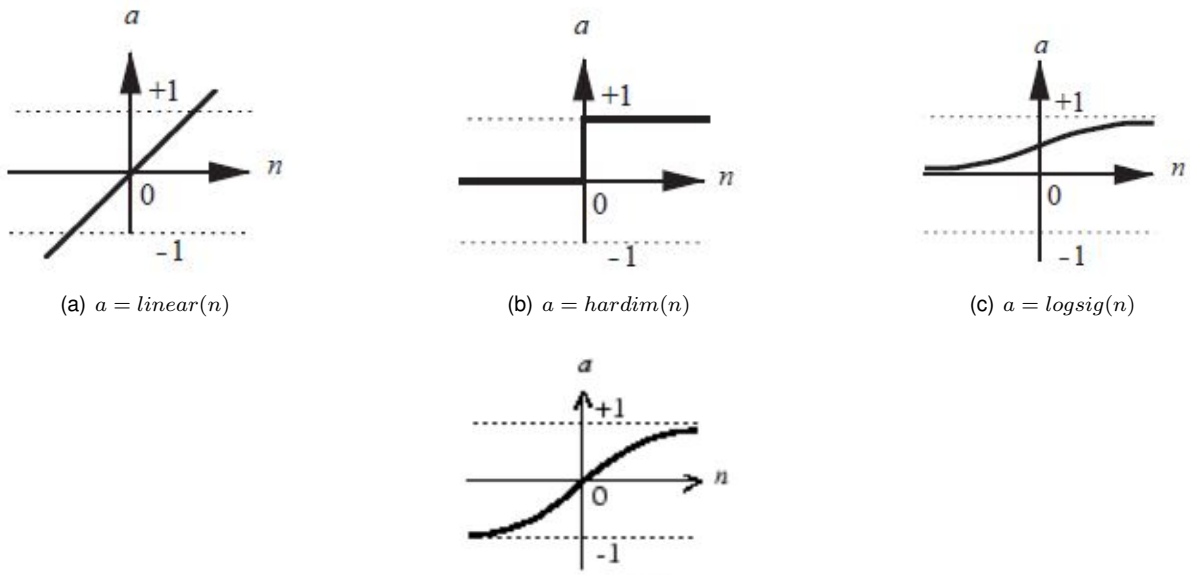

(a) Biological Neuron

(b) Artificial Neuron

Figure 3.2: Biological Neurons vs. Artificial Neuron, [26].

Comparing both biological and artificial neural networks (Figure 3.2) we can observe two key similarities between them. First, the building blocks in each case are simple computational devices and highly interconnected. And for second, the function of the network is, in both cases, determined by the connections between neurons. Their structures have also characteristics that can be considered similar. The Biological Neuron, who has been taken as inspiration for this model, can be divided in three main parts, the Dendrites, a tree-shaped network of nerve fibers that receive from other neurons and conducts the electric signal to the Cell Body, the Axon, a single long fiber that conducts the signal from the Cell Body to other Neurons, and the Cell Body, who sums and integrate the synaptic information before sends it through the axon in addition to his capacity to perform a variety of biochemical processes in order to keep the neuron functioning properly. The point of connection between two different neurons is called Synapse. Depending on the arrangement of the different neurons and their strengths of each synapse the neural network will perform different functions. The Artificial Neurons are composed by a Sum Block

connected to a Transfer Function. The inputs of the Sum Block are the Input Vector multiplied by the correspondent Weight Vector and the Bias. After the sum the result, called Net Input, goes into a Transfer Function, and the result produced by it is called the Neuron Output. The Transfer Function depends on the specification of the problem that the neuron is trying to resolve and can be Linear or Non-Linear (Figure 3.3). Relating with the biological model we can correspond the Weights Vector to the strength of the synapses, to the set composed by the summation and the transfer function we can correspond to the cell body, and the result output from the artificial neural network can be equated to the electrical signal on the axon.




(a) $a = linear(n)$ (b) $a = hardim(n)$ (c) $a = logsig(n)$

(d) $a = tansig(n)$

Figure 3.3: Example of Transfer Functions, [26].

Like a brain or a biological neural network, an ANN is composed by a lot of Neurons and the way that these Neurons are connected between them defines the type of architecture we are using. Usually an ANN is composed by different Layers, a group of neurons receiving the same inputs (Figure 3.4). In this case, each neuron has his Weight Vector and produce a determined Output. Simplifying the mathematical notation and gathering all weights, bias and outputs, each layer can be represented by a Weight Matrix, a Bias Vector and an Output Vector. All the basic elements can be represented by a mathematical expression. The Net Input for the $i$ neuron of the $k+1$ layer can be expressed by:

$$n^{k+1}(i) = \sum_{j=1}^{S^k} (W^{k+1}(i,j) \times a^k(j) + b^{k+1}(i)), \quad k = 0, 1, 2, ..., M-1 \tag{3.2}$$

where $W^{k+1}(i,j)$ is the Weight Vector for the that neuron, $a^k(j)$ is the input from the output of the previous layer and $b^{k+1}(i)$ is the Bias correspondent of that neuron.

When $k = 0$ the input of that neuron is equal to the input of the Neural Network. The Neuron Out-

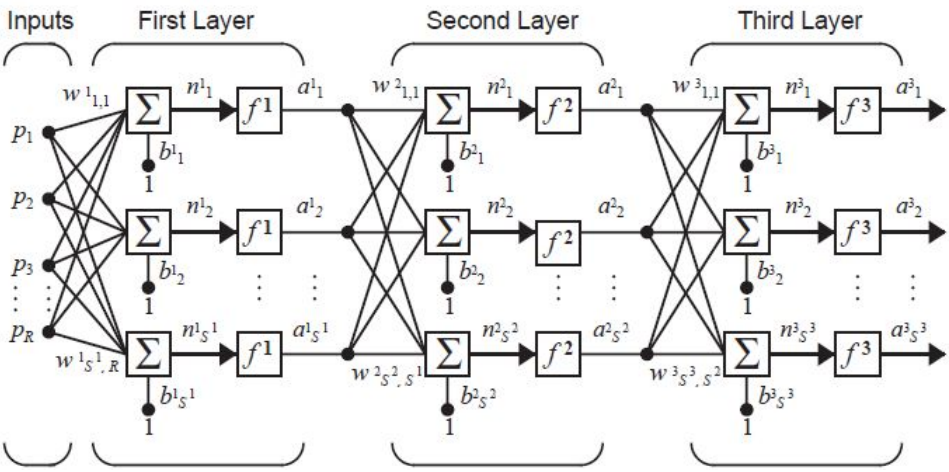



Figure 3.4: Example of ANN Network composed by 3 layers and $S_1$, $S_2$ and $S_3$ neurons respectively, [26].

put for the same neuron and layer can be expressed by:

$$a^{k+1}(i) = f^{k+1}(n^{k+1}(i)), \quad k = 0, 1, 2, ..., M-1 \tag{3.3}$$

where $f^{k+1}(...)$ is the Transfer Function for the $k+1$ layer.

The layer that is responsible to produce the problem's output is the Output Layer. The remaining network's layers are considered the Hidden Layers. The number of elements in an ANN (i.e. number of inputs or outputs, number of layers, number of neurons per layer) are quite arbitrary but must obey some rules. The number of inputs and outputs of the network are defined by the problem specifications, the number of inputs of the network is equal to the number of external variables to be used as inputs and the number of neurons in the Output Layer is equal to the number of outputs of the problem. The number of neurons in the remaining layers will only influence the complexity of the network and this will depend on the problem taken, but the number of neurons of a layer must be equal to the number of inputs of the next layer.

Concerning the influence of the inputs on the outputs of the network we can classify the network as a Static or Dynamic Network. In the first one the output is calculated directly from the current inserted input. In a Dynamic Network the output depends on the current and previous inputs but also outputs or states of the Network since it uses differential equations to calculate the network output. As in the number of ANN elements the choice of Transfer Function in all layers or inside the same layer is flexible and can be chosen one single Transfer Function for all the network or a combination of different Transfer Functions as if they were a partition network of the principal ANN (each partition receive the same inputs and calculate some of the outputs).

After establishing the ANN architecture, what gives the network its adaptive capacity, equivalent to a

biological neural network, is the chosen Learning Process. This process can be classified as Supervised or Unsupervised. In the first case the network receive a set of examples that reflects the expected behavior of the network, composed by an Input Vector and a Target Vector, and modify the weights values through an iterative method comparing the network response to the input received with the target expected, always with the objective of approaching these two vectors. In Unsupervised Learning algorithms, the network weights are corrected based on the inputs delivered to the network using clustering techniques useful on Classification problems.

There are different network architectures and Learning Processes that we can use. In this work the choices fell on a Feedforward Network for the architecture and for the Learning Process the Levenberg-Marquardt Backpropagation Algorithm.

The Feedforward Network is a Multilayer Neural Network composed by two layers, one Hidden Layer and one Output Layer. The transfer functions used in this network are Tan-sigmoid function in the Hidden Layer and the linear function in the Output Layer.

The training algorithm used in this ANN is the Levenberg-Marquardt Algorithm. This algorithm is an variation on Backpropagation Learning Algorithm which is an Supervised Learning Algorithm [27]. The Backpropagation algorithm starts with the Propagation of Input through the network with the equation 3.3. After that the algorithm calculate and propagates the sensitivities back using the next recurrence relation:

$$\underline{\delta}^M = -\dot{F}^M(\underline{n}^M)(\underline{t}_q - \underline{a}_q) \tag{3.4}$$

$$\underline{\delta}^k = \dot{F}^k(\underline{n}^k)W^{k+1^T}\underline{\delta}^{k+1}, \quad k = 0, 1, 2, ..., M-1 \tag{3.5}$$

where

$$\dot{F}(\underline{n}^k) = \begin{bmatrix} \dot{f}^k(n_1^k) & 0 & \cdots & 0 \\ 0 & \dot{f}^k(n_2^k) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \dot{f}^k(n_{S_k}^k) \end{bmatrix} \tag{3.6}$$

is the matrix composed by the derivatives of the transfer functions, and $\underline{t}_q$ and $\underline{a}_q$ are, respectively, the vector of the desired targets and the vector of the output resulting from equation 3.3. The recurrence relation always initialize at the final layer using the equation 3.4 and propagates the sensitivities through equation 3.5. The last step of each iteration is the update of the weights and offsets of the networks. For that, and using the chain rule, we can obtain the update values for the weights and bias of the network through the equations:

$$\Delta\omega^k(i,j) = -\alpha\frac{\partial\hat{V}}{\partial\omega^k(i,j)} = -\alpha\partial^k(i)a^{k-1}(j) \tag{3.7}$$

$$\Delta b^k(i) = -\alpha\frac{\partial\hat{V}}{\partial b^k(i)} = -\alpha\partial^k(i) \tag{3.8}$$

where $\alpha$ is the learning rate and $\hat{V}$ is the performance index. This process is repeated with different input vectors, until the performance error is small enough or the number of iterations reached the maximum number. The Levenberg-Marquardt Algorithm, in addition to having derived from Backpropagation Algorithm, was based on the Newton Method:

$$\Delta \underline{x} = -[\nabla^2 V(\underline{x})]^{-1} \nabla V(\underline{x}) \tag{3.9}$$

where $V(\underline{x})$ is the function we want to minimize with respect the parameter vector $\underline{x}$, and $\nabla^2 V(\underline{x})$ and $\nabla V(\underline{x})$ are respectively the Hessian Matrix and the Gradient of that function. In this specific case is assumed that $V(\underline{x})$ is the sum of squares function:

$$V(\underline{x}) = \sum_{i=1}^{N} e_i^2(\underline{x}) = e^T(\underline{x}) e(\underline{x}) \tag{3.10}$$

where $N = Q \times S_M$ and $e(\underline{x})$ is the error associated to the vector:

$$\underline{x} = [w^1(1,1) \quad w^1(1,2)...w^1(S_1,R) \quad b^1(1)...b^1(S_1) \quad w^2(1,1)...b^M(S_M)]^T \tag{3.11}$$

$$e_q = t_q - a_q^M \tag{3.12}$$

The first step of the Levenberg-Marquardt algorithm is, like the previous one, propagate the input result and compute $V(\underline{x})$. After this step, and instead of compute the sensitivities, the algorithm computes the Jacobian Matrix:

$$J(\underline{x}) = \begin{bmatrix} \frac{\partial e_1(\underline{x})}{\partial x_1} & \frac{\partial e_1(\underline{x})}{\partial x_2} & \cdots & \frac{\partial e_1(\underline{x})}{\partial x_n} \\ \frac{\partial e_2(\underline{x})}{\partial x_1} & \frac{\partial e_2(\underline{x})}{\partial x_2} & \cdots & \frac{\partial e_2(\underline{x})}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial e_N(\underline{x})}{\partial x_1} & \frac{\partial e_N(\underline{x})}{\partial x_2} & \cdots & \frac{\partial e_N(\underline{x})}{\partial x_n} \end{bmatrix} \tag{3.13}$$

which is composed by the derivatives of the errors. The last step is the update of the $\underline{x}$ vector through the equation:

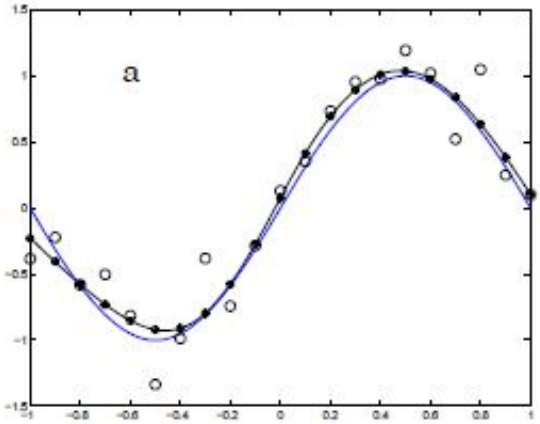
$$\Delta \underline{x} = [J^T(\underline{x}) J(\underline{x}) + \mu I]^{-1} J^T(\underline{x}) e(\underline{x}) \tag{3.14}$$

where $\mu$ and $\beta$ are factors used to control the learning rate and which values are equal to $0.001$ for $\mu$ and $10$ for $\beta$. In the end of these steps is time to compare the sum of square errors $x + \Delta \underline{x}$ with $V(\underline{x})$. If the new sum is bigger than the sum computed in the beginning of the iteration, then $\mu$ is multiplied by $\beta$ and we repeat the computation of $\Delta \underline{x}$. If the new sum is smaller, then $\mu$ is divided by $\beta$ and the process returns to the beginning of the algorithm. The learning process is repeated until the gradient $\nabla V(\underline{x})$ is smaller than some predetermined value, the sum of square errors reduced to a predetermined value, the algorithm reaches the number maximum of iterations or the algorithm reaches the maximum

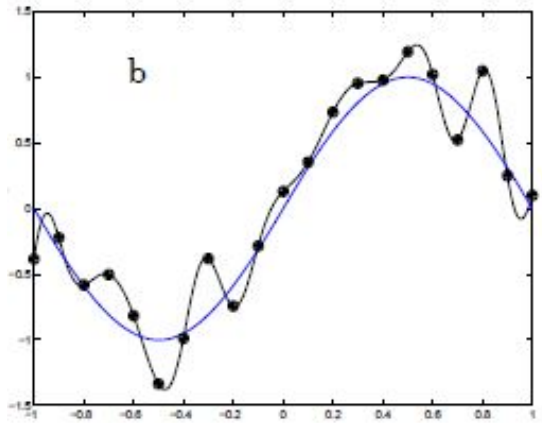
of validation checks. The gradient can be calculated by:

$$\nabla V(\underline{x}) = J^T(\underline{x})\underline{e}(\underline{x}) \tag{3.15}$$

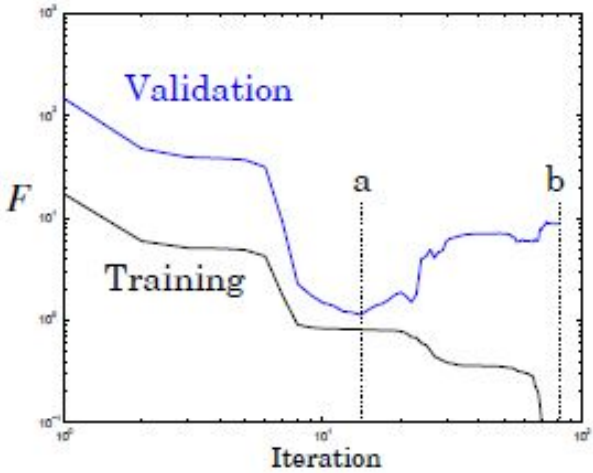
The validations checks belongs to a method of early stopping called Cross-Validation. This method is used to prevent network overfitting and an increasing of the network complexity. In this method the data available to the model is divided in three sets, the Training Set responsible for training the network, the Test Set responsible for giving us an indication of how the network will perform in the future, and the Validation Set used in this method. This set gives the information of what is happening to the network function between the training points. His error on the validation set is monitored and when it goes up several iterations the training is stopped, and the weights and bias that produced the minimum error on the validation set are used in the final prediction. The Figure 3.5 shows an example of an early stopping process using cross validation.




(a) Minimum Validation Error Case




(b) Last Iteration Case




(c) Evolution of the Validation and Training Sum Squared Errors, $F$, throughout each iteration

Figure 3.5: Example of Early Stopping using Cross Validation in the process of fitting the network response in the blue line function, [26].

## 3.3 K-Nearest Neighbors (KNN)

The K-Nearest Neighbor (KNN) Model is, like the ANN method, an Artificial Intelligence method, being also considered a System of Pattern Recognition [28][29]. This type of systems are able to identify a random object knowing previously the nature of one set of similar objects.

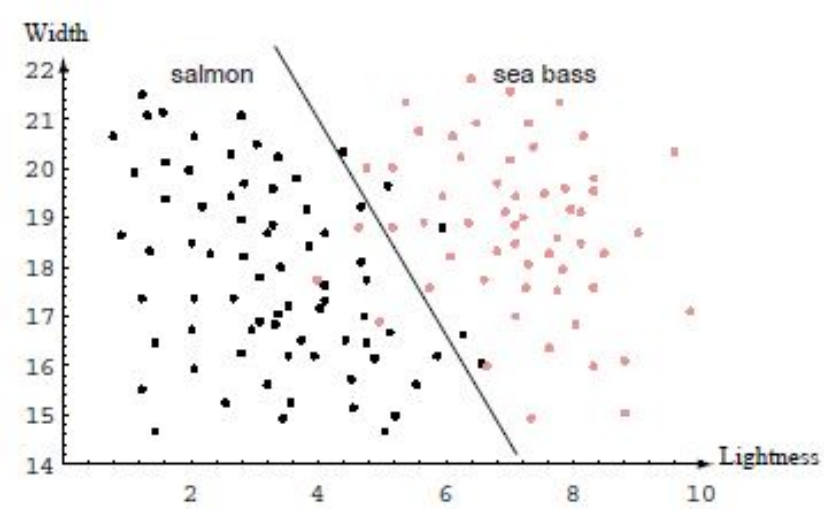



Figure 3.6: Example of KNN Classification Problem. Classification of two fish types based on two parameters (Width and Lightness). An unknown point above the black line has more probability of being a Sea Bass. Below of her the probability of being Salmon is bigger, [29].

This method started to be used in Classification problems. The main objective of these problems is to assign a classification to a chosen point based on the class of the training set points. This process can be compared to the human sensory capacity like recognizing a face, identify our accessories in our bag by feeling their shape, understanding different words in different languages by listening or decide if an apple is ripe only by its smell.

This Model can also be classified according to his training algorithm and characteristics. Just like the ANN, the KNN Learning Process is classified as a Supervised Learning Process because it attributes to each input object an target value. These systems can be divided into Parametric or Nonparametric Techniques. In the first case is known the expression for Probability Density or other Discriminant Functions (e.g. Maximum Likelihood Method). The Nonparametric Techniques does not possess that information (Discriminant Functions) *a priori*. In this case, the estimation problem is formulated with great generality in a space of non-negative functions, thus not restricting the probability density function to belong to a restricted class of functions that depend on a parameter vector. It is in this class of techniques that the KNN Model is inserted.

The KNN method is an application of the Maximum a Posteriori Probability (MAP) Classification with the Probability Density Functions obtained by the Parzen Method with Adaptive Window. Mathematically lets consider $X$ a set composed by $N$ classified training patterns and $X^i$ a subset of $X$ generated

by the $i$-th class. By the MAP classifier, the classification of a pattern $x$ is given by:

$$\hat{\omega} = \arg\max_{\omega} \hat{p}(x/\omega)\hat{P}(\omega) \tag{3.16}$$

where $\hat{p}(x/\omega)$ is the probability density function estimation and $\hat{P}(\omega)$ is the *a priori* probability estimation for class $\omega$. The estimation $\hat{p}(x/\omega)$ can be obtain through the Parzen Method. The window used in point $x$ is chosen so that only the $k$ nearest patterns to $x$ are inside the same window and thus contributing to obtain the probability density functions:

$$w(x - \tilde{x}) = \begin{cases} c, & \tilde{x} \in V_\delta(x) \\ 0, & \text{otherwise} \end{cases} \tag{3.17}$$

where $c$ is a normalization constant and:

$$V_\delta(x) = \{\tilde{x} : \|\tilde{x} - x\| < \delta\} \tag{3.18}$$

is the circle region of radius $\delta$ centered in $x$, composed of points whose distance to $x$ is less than $\delta$. As the aim is to ensure that the window possess only the k-nearest training patterns to $x$, we can get this by defining the value of $\delta$ as the distance of the $(k+1)$-th nearest training pattern to $x$. From the probability calculus and from the MAP classification we know that:

$$\hat{P}(\omega_i) = \frac{N_i}{N}, \quad i = 1, \ldots, t_c \tag{3.19}$$

$$\hat{p}(x) = \frac{1}{N} \sum_{\tilde{x} \in X} w(x - \tilde{x}) \tag{3.20}$$

where $t_c$ represents the number of classes. By combining the previous equation with 3.17 we can rewrite 3.16:

$$\hat{p}(x/\omega_i)\hat{P}(\omega_i) = \frac{N_i}{N}\frac{1}{N_i} \sum_{\tilde{x} \in X^i \cap V_\delta(x)} w(x - \tilde{x}) = \frac{ck_i}{N} \tag{3.21}$$

where $N_i$ is the number of elements of $X^i$ and $k_i$ is the number of training patterns of the $i$-th class belonging to $V_\delta(x)$. The sum of all $k_i$ is equal to $k$. The class that maximizes 3.21 is obtained by the equation:

$$\hat{\omega} = \omega_p : p = \arg\max_{i} k_i \tag{3.22}$$

which returns the $k_i$ that maximizes the equation 3.21, which is the class most represented in the subset of the k-nearest neighbors. In short, the KNN Classification Algorithm can be explained in three steps. First, we need to calculate the distances between the pattern $x \in \mathbb{R}^n$ (the point that we want to classify) and the other training patterns (Figure 3.7). This calculation is obtain through a Distance Function. The

most used for continuous variables are:

$$\text{Euclidean} \quad \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{3.23}$$

$$\text{Manhattan} \quad \sum_{i=1}^{k}|x_i - y_i| \tag{3.24}$$

$$\text{Minkowski} \quad (\sum_{i=1}^{k}(|x_i - y_i|)^q)^{1/q} \tag{3.25}$$

After that, we select the k training patterns nearest to $x$. Finally, we determine the most present class among all k-Nearest Neighbors. The classification for $x$ is the same class previously determined.

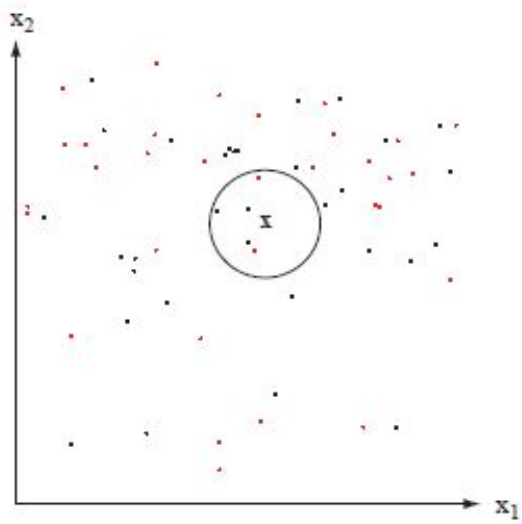



Figure 3.7: Example of KNN Classification Algorithm. In this case the training patterns are arranged according to two parameters and the algorithm search for the 5 Nearest Neighbors to $x$, [29].

The KNN Method can also be used to solve Regression Problems. In this case, instead of classify, the main objective is to determine the numerical value of a variable of a unknown case, which means, each training pattern is not associated to a class but to a numerical value of a specific variable. To solve a problem like this, the algorithm of KNN Regression is in all similar to the one of KNN Classification except for the last steps. It starts the same way calculating the distance between all the training patterns and the point to be studied, $x$. After this, and after discovering the k-nearest patterns, the variable numerical value of $x$ is equal to the average of all the variable numerical values of the k-Nearest Neighbors.

This is the simplest algorithm for KNN Regression but there are other variations for this Method [30]. In the next chapter will be explained how this model was applied in this case study and the variations that were tested in this work.

# Chapter 4

# Preparation of Case Study Simulations

For a better comparison between models, both ANN and KNN were applied to the same Case Study. In this work we are analyzing and trying to predict one day of Solar Irradiance in the City of Lisbon. In the next Sections we will explain in detail the characteristics of the data used in this study. We will also analyze how the ANN Model and the KNN Model were applied to this case study and show how the main parameters have been chosen for the best performance of the models forecasting.

## 4.1 Data

As the main objective of this thesis is to predict with the best performance the Solar Irradiance of one day, the data used for this purpose is the Solar Irradiance registered in Lisbon for 365 days, or one year. In this Case Study, beside knowing that the Solar Irradiance is, directly and indirectly, influenced by other meteorological parameters (Date and Time, Precipitation, Cloud Cover, Air Pollution, etc.), these parameters were not considered and were not used in any test or simulation during this work.

The values of Solar Irradiance can be obtained from Lisboa_ano.txt document to be easy to access in the developed algorithms. This document includes a total of 8670 records of Solar Irradiance, one for each hour of the year. In each test, or simulation, this group of values are stored in a matrix that organizes the values by hour of the day, by day of the month and by month of the year for an easier access during the study.

After creating the Data Matrix, to gather the training matrix each Model copies the values registered in the past from the data matrix and stores them on a new matrix. Depending on the model it will store them in a different way due to the requirements of each model. This aspect will be demonstrated in more detail in the next section.

## 4.2 Application of ANN to the Case Study

To apply the ANN model to this case study we will need 3 matrices based on the Data Matrix, a Training Matrix, a Target Vector and an Input Vector. These matrices are created and filled for each hour forecast, which means that different predictions have different matrices.

The Training Matrix is the Matrix that stores all the information of the past hours Irradiance. Each column of the matrix contains the past information of a certain hour. The number of rows depends on the information we want to store about the same hour. For example, if we used $N$ hours to train a network and the last 3 records of Solar Irradiance before hour x ($SI_x$), the Training Matrix has $N-3$ columns and 3 rows storing the records from $SI_{x-3}$ to $SI_{x-1}$.

The Target Vector is simply the vector that stores the real value of Irradiance for one specific hour. This vector has only one row, due to the fact that it saves only one kind of information, but it has exactly the same number of columns of the Training Matrix. The number of columns is mandatory being the same in both matrices to make the connection between both information (the past information and the real information). Considering the last example, the Target Vector has $N-3$ entries always storing the record $SI_x$. The Figure 4.1 shows the algorithm mentioned before. The Blue Squares represent the records stored in the Training Matrix and the Red Square represent the record saved in the Target Vector for each cycle.

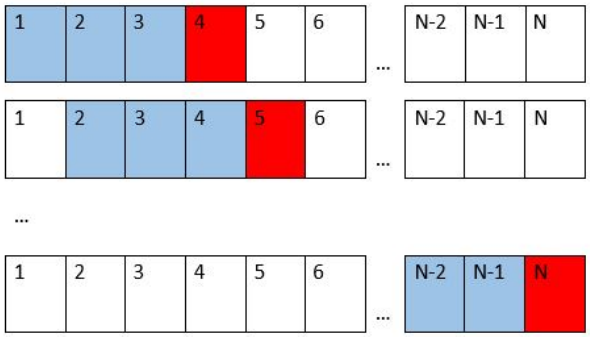



Figure 4.1: Storing algorithm in the ANN Model.

Finally, the Input Vector is the vector used to predict the next value of Solar Irradiance. The dimension of this vector has to be equal to the number of rows of the Training Matrix and it includes the records right previous to the hour supposed to predict. Using the last example, if we want to predict hour p ($SI_p$) then the Input Vector is composed of $SI_{p-3}$, $SI_{p-2}$ and $SI_{p-1}$.

After creating both Matrices the next step on ANN Model will be to train the Network and predict the next value of Irradiance. For each hour predicted the program creates and takes the Training Matrix and the Target Vector and fits the network for the records present in both matrices. When this fitting

is completed the program will predict the next value applying the Input Vector into the network and the predicted value is the Output given by the network. This process is repeated as many times as the number of desired predicted hours. This algorithm is represented in Figure 4.2.

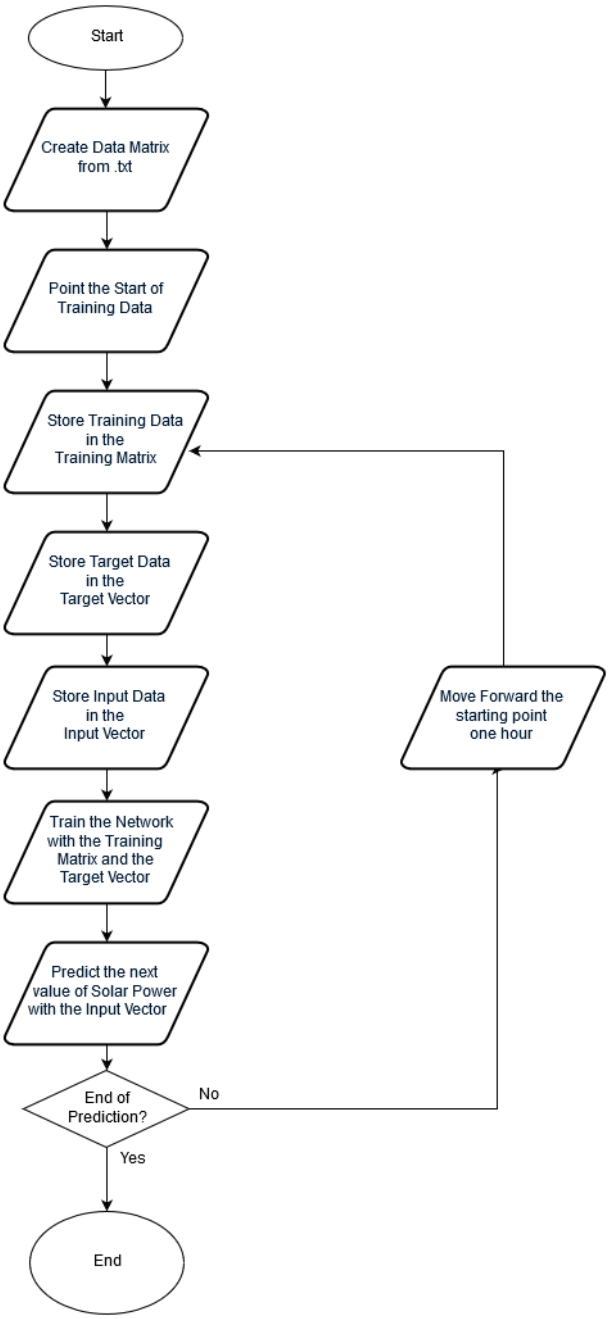



Figure 4.2: ANN Algorithm used in this work.

In order to create this algorithm in the Matlab® software we used some of the tools already existing in the Matlab® Toolboxes, namely, in the Neural Network Toolbox. The most used tools in this work were the *feedforwardnet* command to create a Feedforward Network, the *train* command to train the network based on the training matrix and the target vector, the *net.trainParam.epochs* to change the number of training epochs (which is the equivalent for the training iterations explained in the previous chapter) and

the *sim* command to predict the next value of Solar Irradiance.

To get the best performance of the model it is necessary to test the main parameters (e.g. hidden layer size, number of epochs, type of input information, number of days to train the network). In the next subsection we will briefly demonstrate all the tests carried out to gather the best parameters.

### 4.2.1 ANN Tests

These tests were taken based on the algorithm previously explained and just changing the parameter under evaluation. The results of these tests can be consulted in Section 5.1.

#### Hidden Layer Size and Epochs

At the beginning of this thesis the first problem to deal with was the best Hidden Layer Size and the best number of Maximum Training Epochs. In order to achieve this, the tests were made under the same conditions and fixing one of the parameters and changing the other. With the fixed Hidden Layer we made the tests with 18 Hidden Layers and when the Epochs were fixed the tests had 1000 epochs. For training the network we used 288 hours (12 days) always to predict the 289$^{th}$, starting the test with the first 288 records of the data file and ending the test after the forecast of 5 days.

#### Type of Information

After deciding what was the best Hidden Layer Size and the best number of Training Epochs the next parameter to test was the kind of information we should use to perform the Irradiance Forecast. Excluding all the meteorologic parameters and trying to find a balanced quantity of information we tested 2 different methods of filling the Training Matrix.

The first one, named Past Records Method, uses past records of solar power and also stores the value of the hour that is going to be predicted. The firsts tests of the ANN Model showed predictions during the night period with values different from zero. That prediction is totally incorrect and this solution, of storing the hour to be predicted (0 to Midnight 12 to Midday) together with the records of past solar power, brought more precision to the prediction.

The second method, named Variation Method, also stores the value of the hour to be predicted but didn't store the records of solar power like the last model. This method only stored the last record of Solar Irradiance and instead of more records it stored the variation between the last 3 hours of Irradiance, ($SI_{p-3} - SI_{p-2}$) and ($SI_{p-2} - SI_{p-1}$).

Since both methods are still better than using just the records of solar power we tested each one to verify which was the best option to store the information in the Training Matrix. This test was made with

one month of records to train the network and the test ends after the forecast of the Solar Irradiance for the next 24 hours.

#### Delays and Training Time

After identifying the best way to organize the information for the forecasts it was time to understand the relation between the past records and the present record of Solar Irradiance. In order to do so, the best way was to use Partial Autocorrelation on the data available in the Lisboa_ano.txt document. The result can be analyzed in Figure 4.3.

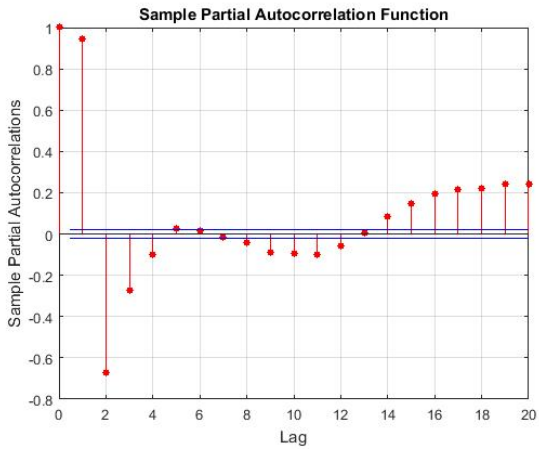
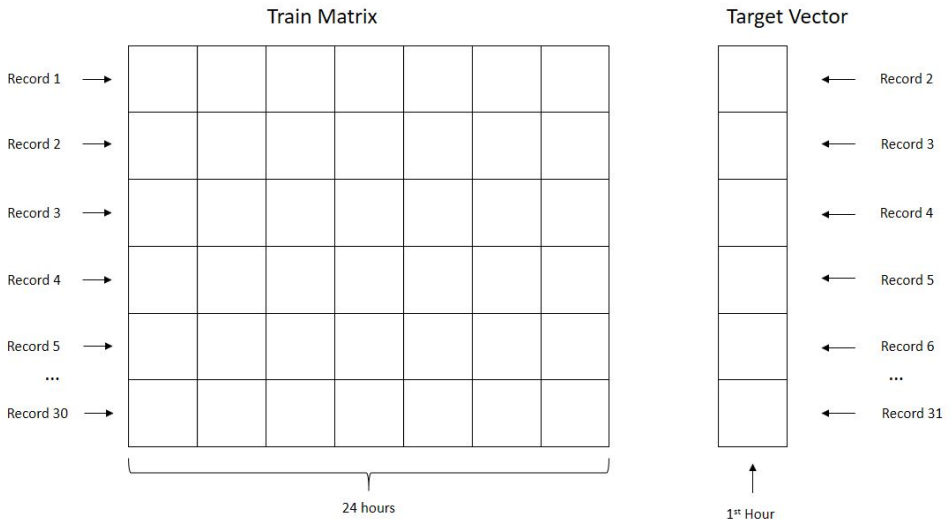



Figure 4.3: Partial Autocorrelation of Irradiance Data.

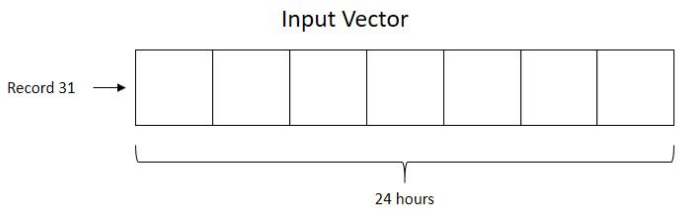
After the examination of the Partial Autocorrelation the model was tested with different time delays (2, 3, 4 and 5 hours delay) to be inserted in Training Matrix and for different training times (3 days, one and two weeks, one month and one month and a half). The time delays will be reflected on the number of rows of the Training Matrix (more delays, more rows) and the training time will be reflected on the number of columns of the same matrix. This test was also made with one month of records to train the network and the test ends after the forecast of the Solar Irradiance for the next 24 hours.

## 4.3 Application of KNN to the Case Study

Like the ANN Model, the KNN needs 3 matrices named Train Matrix, Target Vector and Input Vector. The purpose of these matrices is the same as the purpose of the matrices in the ANN Model but the way that the information is stored is different. The Train Matrix stores the records of Solar Irradiance of different days in each row (by default was used records of 24 hours length), the Target Vector stores the Solar Irradiance record registered after the records of the Train Matrix, and the Input Vector stores the record of Irradiance of the last hours before the hour to be predicted. For example, if the Train Matrix is composed of 30 records of 24 hours length then the Target Vector length is equal to 30 and the Input Vector is also composed of one record of 24 hours. The figures in figure 4.4 illustrate the above explanation for the case of records of 24 hours in each row of the Train Matrix.




(a) KNN's Train Matrix and Target Vector




(b) KNN's Input Vector

Figure 4.4: Matrices used in KNN Model to forecast the first hour of the 32$^{nd}$ day.

The KNN algorithm has some resemblances with the ANN algorithm, only differing in the training model process. After creating both matrices the program calculates the distance between each row of the Train Matrix and the Input Vector. Since every distance functions determine one distance between two points and the performance of the forecast is not affected by the chosen function, the distance function used to determine the k nearest points was the Euclidean Distance. After analyzing which one of the records was the k nearest neighbors of the Input Vector the predicted hour was defined by the average value between the k records of the Target Vector correspondent of the k nearest of the Train Matrix. The flowchart illustrating this process is represented in Figure 4.5.

Since the version of Matlab® used in this work does not have the tools for KNN Regression, but only for KNN Classification, the results using those tools were not quite correct. To solve this problem we created an algorithm using simpler tools to calculate the distance between points and the average of the target records.

As in ANN case, before testing the real performance of the KNN Model, the test of the main parameters of this model was required. In the next subsection there is a brief explanation of these tests.

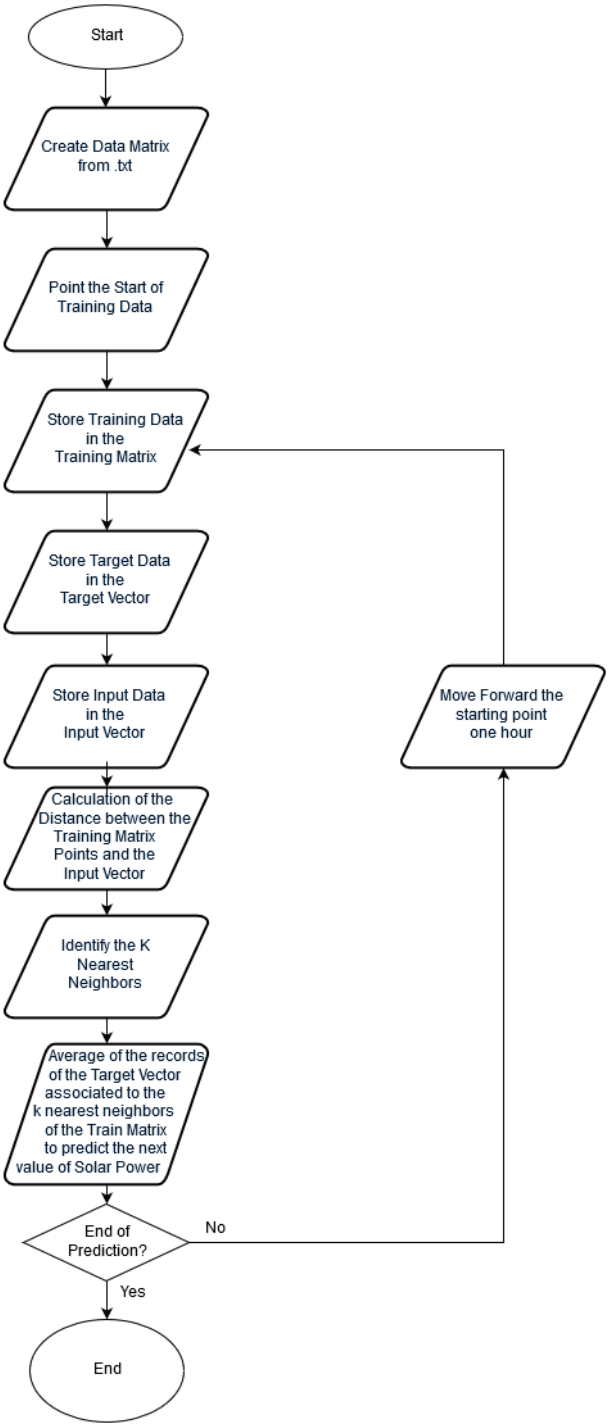



Figure 4.5: KNN Algorithm used in this work.

### 4.3.1 KNN Tests

Like the ANN Tests, the KNN tests were carried out based on the previous algorithm but changing the parameters under scrutiny. The first parameter to be tested was the most important and the one which gave the name to this model, the number of K Neighbors. The results of KNN Tests can be consulted in section 5.2.

**K Parameter and Distance Method**

The first approaches to this model (tests for predictions of different days) revealed that there was not an ideal k number for both predictions This is a problem mentioned in several articles. For different training datasets the ideal k number is also different. To overcame this problem two approaches were tested, a Weighted Distance Calculation and an Ensemble Learning Approach adapted to this regression case.

The first approach is very common and gives priority to the neighbors that are nearer to the input vector. To forecast the next value of Solar Irradiance ($SI_{d+1}$) with a Weighted Distance Calculation we used the following equation:

$$SI_{d+1} = \frac{1}{\alpha_1 + ... + \alpha_k} \sum_{l=1}^{l=k} \alpha_l \cdot T_l \tag{4.1}$$

where the $\alpha_1$,...,$\alpha_k$ are the weights for the different distances and $T_1$,...,$T_k$ are the targets correspondent of the k nearest neighbors.

In what concerns the calculation of the weights, these were based on the distance between the training points and the Input Vector. The nearer the points the bigger the $\alpha_l$ are. The equation used to the acquisition of the weights is:

$$\alpha_l = \frac{d_w(SI_{IV}, SI_k) - d_w(SI_{IV}, SI_l)}{d_w(SI_{IV}, SI_k) - d_w(SI_{IV}, SI_1)} \tag{4.2}$$

where $d_w = (a, b)$ is the calculated distance between the points $a$ and $b$, $SI_{IV}$ is the point stored in the Input Vector, $SI_1$ is the k nearest neighbor nearer to the Input Vector, $SI_k$ is the k nearest neighbor further of the Input Vector and $SI_l$ is the k nearest neighbor correspondent of $\alpha_l$.

The second method, or the Ensembling Learning Approach (ELA), takes the most part of the predictions of Solar Irradiance (from 1-Nearest Neighbor to $\sqrt{N}$-Nearest Neighbor predictions, where $N$ is equal to the number of points in the training matrix) to make a new prediction. The algorithm is based on the article [30] in which the author uses this approach in a KNN Classifier. In this work, this approach was adapted for a KNN Regression. The program predicts the value of Irradiance for a certain hour using different numbers of K Neighbors (1,2,...,$\sqrt{N}$). After that the predicted value is equal to the weighted average of all the predictions. The Weighting Function is an Inverted Logarithmic Function, similar to the one proposed in the article

$$w(k) = \frac{1}{\log_2(1 + k)} \tag{4.3}$$

in which $k$ is the number of the nearest neighbors.

This test includes testing the performance of the model with or without the Weighted Distance Function and the Ensemble Learning approach with or without the Weighted Distance Function for 3 days with different weather profiles, cloudy days and clear sky days (1$^{st}$ and 23$^{rd}$ of February and 8$^{th}$ of April). The choice of these days was made with the intention of trying to cover the maximum of situations that we could encounter during the work. Like some ANN tests this test was made with one month

of records to train the model and the test ends after the forecast of the Solar Irradiance for the next 24 hours.

**Record Lenght**

After testing the K parameter the next step is, like the ANN Model, to test the record length used in each record of the Training Matrix.

In the first tests, the train matrix records used 24 hours to train the model. Now this test will predict the same days as before but using different amount of days to train the model. With records bigger then 24 hours the Training Matrix will also increase the size of the columns. For records with 48 hours (2 days) the matrix gets 48 columns, for records with 72 hours (3 days) the matrix gets 72 columns and so on. To prevent the decreasing of the number of points used in the training, the data in the Training Matrix was stored in a different way. In Figure 4.6 we have an example of storing the information in the train matrix for records with 42 hours. Once again, this test was made with one month of records to train the model and the test ends after the forecast of the Solar Irradiance for the next 24 hours.

Hours

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Record 1 | 1 | ... | 23 | 24 | 25 | ... | 48 |
| Record 2 | 25 | ... | 47 | 48 | 49 | ... | 72 |
| Record 3 | 49 | ... | 71 | 72 | 73 | ... | 96 |

Figure 4.6: Information storing process of the Train Matrix for a 2 days record example in the KNN Model.

**Number of Records**

The last test before examining the best performance of the KNN Model is the number of records used to train the model. In the previous tests we took by default 30 days to train, which means that the Train Matrix has 30 rows of dimension. This test will analyze the performance of the model predicting 3 different days using different amounts of days in the Train Matrix (from 5 days to 60).

In this case it was not possible to predict the same days as in the previous tests ($1^{st}$ of February, $23^{rd}$ of February and $8^{th}$ of April) because it was not possible to train the model with 60 days (60 records) and predict the $1^{st}$ of February since the data of the Lisboa_ano.txt document starts on $1^{st}$ of January. For that reason, we changed the days to predict ($3^{rd}$ of March, $8^{th}$ of April and $1^{st}$ of December) but we kept a similar weather profile by choosing days with Persistence errors similar to the previous days.

This time, this test takes different amount of days records of Solar Irradiance to train the model but the target is the same (the simulation only ends after the forecast of the Solar Irradiance for the next 24

hours).

## 4.4 Performance Evaluation

To make the analysis as rigorous as possible the tests must be carried out in the same conditions whenever possible. All the tests were taken in the same computer in order not to jeopardize the analysis of the performances. Each test was repeated 10 times. This was due to the fact that some results were not equal every time a simulation runs. In both methods the time of simulation has small variations related with the speed of computational processes. In ANN case, as the initial network's weights are randomly chosen the value of the forecast for which the model converges is not always the same. In the KNN case the performance error does not change because the points selected for the forecast remain the same in different simulations. Therefore, to obtain an accurate result we made an average between the 10 results of each test repetition.

In order to obtain the performance error of each forecast and in order to analyze every parameter and case in a better way it was necessary to use some mathematical tools, which are very common in this kind of studies:

- Root Mean Square Error: $RMSE = \sqrt{\frac{\sum_{t=1}^{N}(z_t - SI_{real})^2}{N}}$
- Mean Absolute Error: $MAE = \frac{\sum_{t=1}^{N}|z_t - SI_{real}|}{N}$
- Mean Absolute Percentage Error: $MAPE = \frac{1}{N}\sum_{t=1}^{N}|\frac{z_t - SI_{real}}{SI_{real}}|$

where $z_t$ is the value of Solar Irradiance predicted in the model, $SI_{real}$ is the real value and $N$ is the total number of predictions. For the night period as we know *a priori* that the forecast for this period is equal to zero then the calculation of the performance error for this period is not considered in every tests. We will considerate the error when the actual Solar Irradiance is superior to 10 $W/m^2$.

The principal evaluation criteria used in this work is the MAPE. Nonetheless we have used other criteria for possible indecisions between different hypothesis. In some cases these 3 types of errors were not enough to decide which was the best case. Then, we calculated the relation between the error of the forecast and the Persistence Model as you can see in the next equation.

$$MAPE_{model} = \frac{MAPE_{pred}}{MAPE_{pers}} \tag{4.4}$$

If the prediction is equal to the result in the Persistence Model then the $MAPE_{model}$ is equal to 1. The lower the $MAPE_{model}$ the better the performance is and the better is comparing to the Persistence Model.

# Chapter 5

# Results

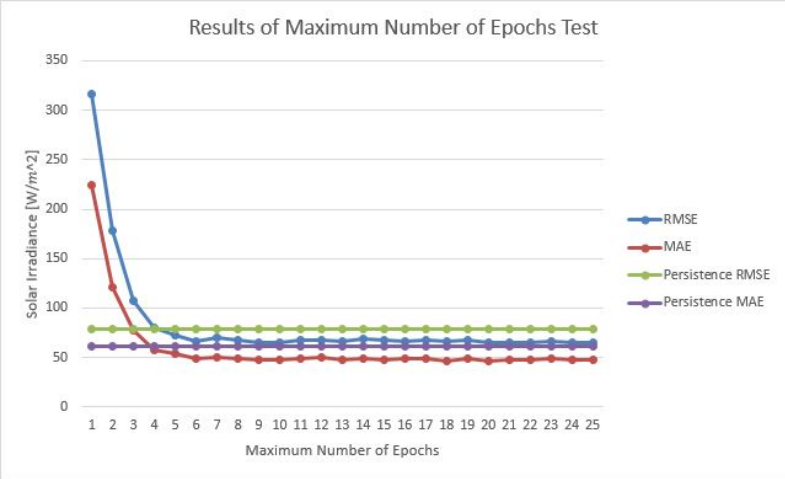
In this chapter the results of the tests explained in the subsections 4.2.1 and 4.3.1 will be presented in sections 5.1 and 5.2. We will also present the reasons for the decisions made after the analysis of these tests. In the last section of this chapter we are going to explain and show the results of the comparative study of both models.

## 5.1 The Results of the ANN Tests

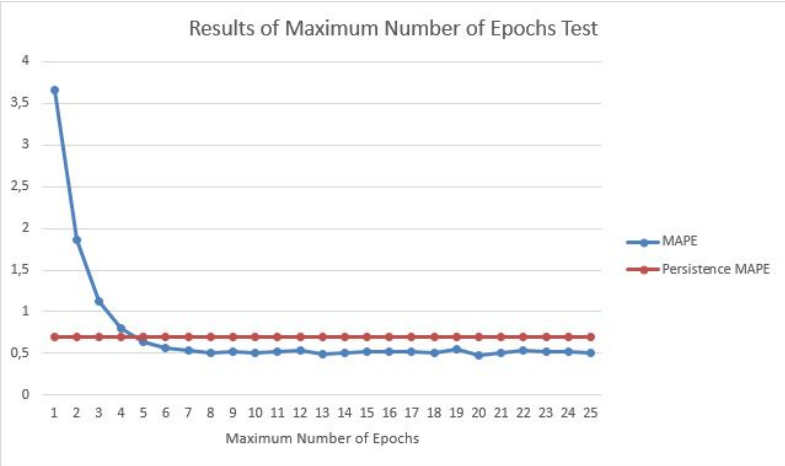
As explained in the subsubsection 4.2.1 the first thing to test was the best number for the Hidden Layer Size and the Maximum Number of Epochs. When we tested the Number of Epoch we fixed the Hidden Layer Size in 18 neurons. When we tested the Hidden Layer Size we fixed the Number of Epoch in 1000. The results of the Number of Epochs can be consulted in Figure 5.1. In the first cases, in which the number of Training Epochs is between 1 and 4, the simulation error is, in general, higher than the Persistence Model due to the fact that the training process doesn't reach the number of validation checks. With more epochs the network is capable of reaching the maximum validation checks more easily improving the simulation error. After analyzing the figure 5.1 it was evident that more epochs give a better prediction, but the time of simulation is also higher than the cases with less epochs. For these reasons it was decided not to change the maximum number of epochs in favor of a better prediction, which is by default equal to 1000 epochs.

For the case of the Hidden Layer Size we can observe in Figure 5.2 that the best case is the one with 4 Hidden Layers, but this is one of the most time-consuming cases. Because of that it is important to find one case that brings a good forecast and that has an acceptable simulation time. The 20 Hidden Layer case, with a MAPE approximately 6% higher than the better case and 22 seconds quicker than the 4 Hidden Layer case, was the size chosen for the Neural Network in this work.

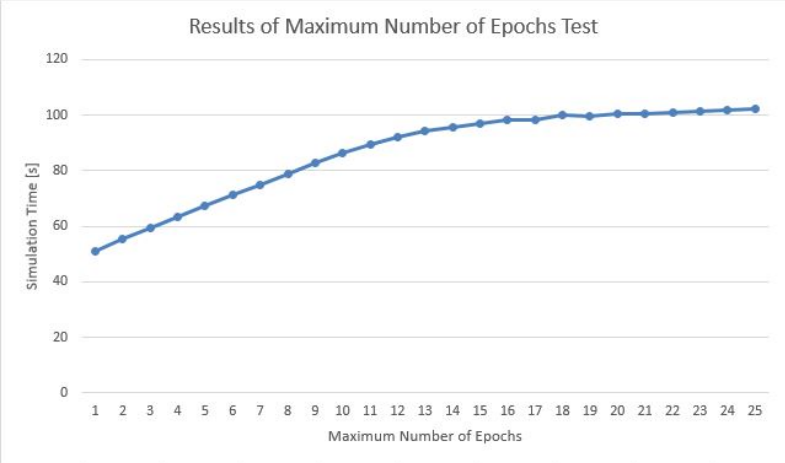
After determining the Number of Training Epochs and the Hidden Layer Size it was time to test the way to store the information to train the network. This test was made with one month of records for train-




(a) RMSE and MAE



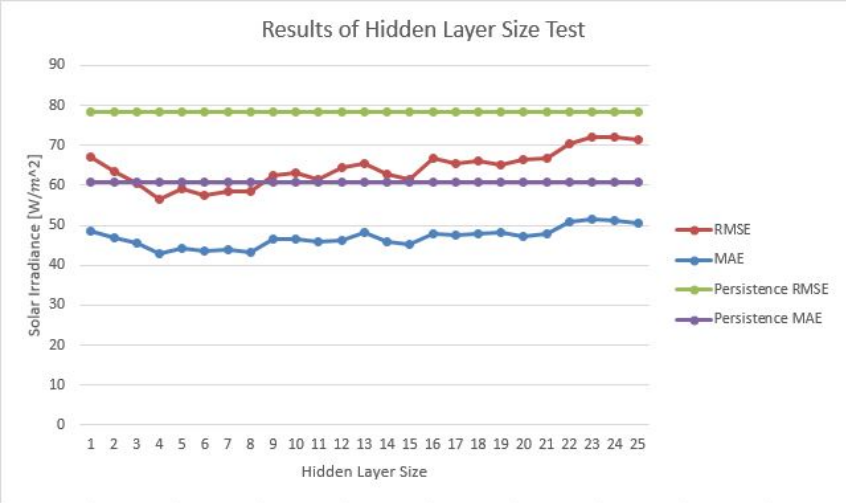

(b) MAPE



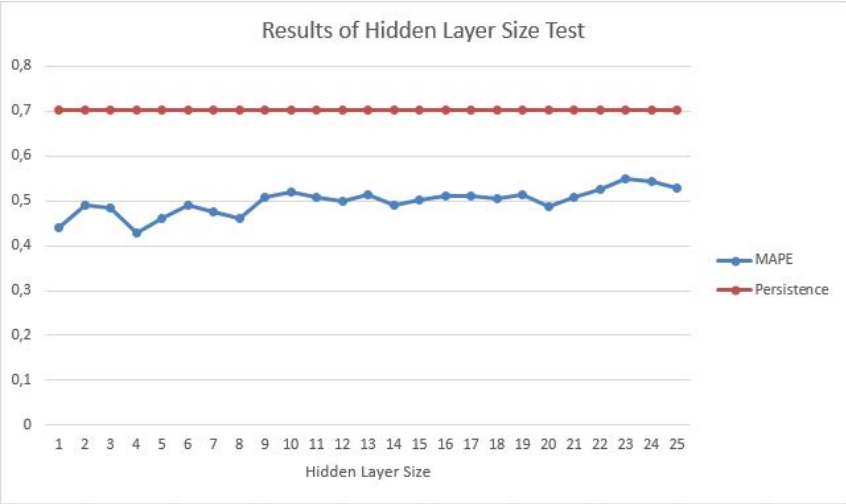

(c) Simulation Time

Figure 5.1: Results of ANN Test to Determine the best Number of Training Epochs.

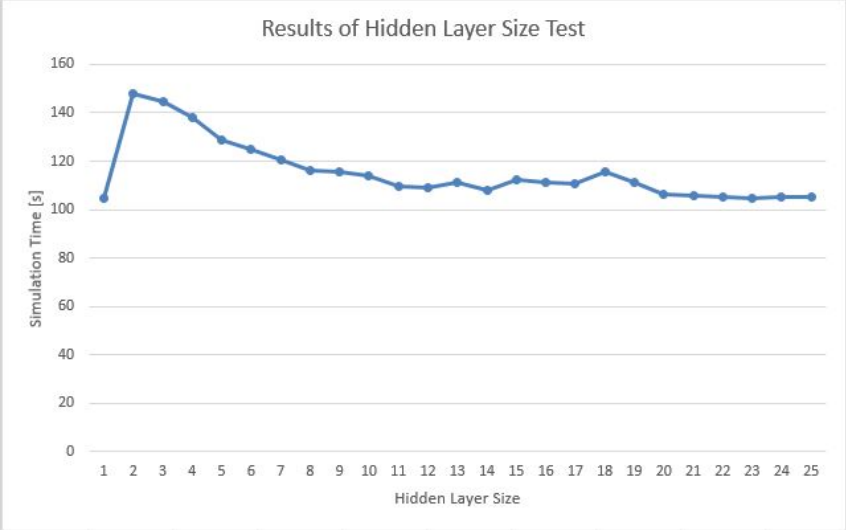
ing and ends after the forecast of the next 24h. The tests were repeated 10 times and made the average of those 10 simulations. The result is represented in Table 5.1. As we can see, although MAPE is worse, the fact that Past Records Method is quicker and both methods have very close results were the decisive factors that led us to adopt this method to store the information in the Train Matrix existent in Data Matrix.




(a) RMSE and MAE




(b) MAPE




(c) Simulation Time

Figure 5.2: Results of ANN Test to Determine the best Hidden Layer Size.

Table 5.1: Results of the ANN Test to Determine the best Storing Method

| **Model** | **RMSE** $[W/m^2]$ | **MAE** $[W/m^2]$ | **MAPE** | **Simulation Time** $[s]$ |
|---|---|---|---|---|
| **Persistence Method** | 111.05 | 104.40 | 0.6351 | |
| **Past Records Method** | 74.58 | 55.33 | 0.3104 | 29.89 |
| **Variation Method** | 74.99 | 56.84 | 0.3026 | 39.80 |

The last parameters test before the final simulations compared not only the amount of delays used in the storage method but also the training time of the network. The results of this test are displayed in the tables 5.2, 5.3, 5.4 and 5.5.

Table 5.2: RMSE of Delay and Training Time Test for ANN

| **RMSE** $[W/m^2]$<br>Pers : 111.05 | | | | | |
|---|---|---|---|---|---|
| | **Delay** | | | | |
| **Training Time** | **2** | **3** | **4** | **5** | **Best** |
| **1 Month and a Half** | 74.68 | 77.93 | 78.47 | 78.21 | 74.68 |
| **1 Month** | 72.76 | 76.89 | 80.73 | 76.03 | 72.76 |
| **2 Weeks** | 77.32 | 86.80 | 84.31 | 82.43 | 77.32 |
| **1 Week** | 104.94 | 118.12 | 106.62 | 106.43 | 104.94 |
| **3 Days** | 265.56 | 224.19 | 257.61 | 245.29 | 224.19 |

Table 5.3: MAE of Delay and Training Time Test for ANN

| **MAE** $[W/m^2]$<br>Pers : 104.40 | | | | | |
|---|---|---|---|---|---|
| | **Delay** | | | | |
| **Training Time** | **2** | **3** | **4** | **5** | **Best** |
| **1 Month and a Half** | 53.17 | 58.66 | 59.49 | 58.39 | 53.17 |
| **1 Month** | 54.15 | 58.28 | 61.05 | 56.80 | 54.15 |
| **2 Weeks** | 64.52 | 72.79 | 65.41 | 63.47 | 63.47 |
| **1 Week** | 87.44 | 99.07 | 89.05 | 85.92 | 85.92 |
| **3 Days** | 205.09 | 171.57 | 191.09 | 180.04 | 171.57 |

The results obtained show that the case with better RMSE is the case with 2 hours delay and 1 month data for training, but the case with 2 hours delay and 1 month and a half of training have better MAE and MAPE. As expected, the fastest simulation is the case with 2 hours delay and 3 days for training the network. This is due to the less amount of information to process in comparison with other cases.

After the ANN Parameters Tests, the final simulations with this model are made with:

- a Storage Method that includes past records of Solar Irradiance and the hour to predict;

Table 5.4: MAPE of Delay and Training Time Test for ANN

| **MAPE** Pers : 0.6351 | | | | | |
|---|---|---|---|---|---|
| | **Delay** | | | | |
| **Training Time** | **2** | **3** | **4** | **5** | **Best** |
| **1 Month and a Half** | 0.2892 | 0.3313 | 0.3466 | 0.3398 | 0.2892 |
| **1 Month** | 0.2971 | 0.3372 | 0.3679 | 0.3233 | 0.2971 |
| **2 Weeks** | 0.3528 | 0.4433 | 0.3544 | 0.3453 | 0.3453 |
| **1 Week** | 0.4331 | 0.5697 | 0.5625 | 0.4921 | 0.4331 |
| **3 Days** | 1.1500 | 0.7844 | 1.0959 | 1.1053 | 0.7844 |

Table 5.5: Simulation Time of Delay and Training Time Test for ANN

| **Simulation Time** $[s]$ | | | | | |
|---|---|---|---|---|---|
| | **Delay** | | | | |
| **Training Time** | **2** | **3** | **4** | **5** | **Best** |
| **1 Month and a Half** | 36.97 | 38.12 | 40.35 | 48.33 | 36.97 |
| **1 Month** | 31.16 | 31.70 | 35.51 | 39.06 | 31.16 |
| **2 Weeks** | 23.83 | 23.43 | 26.07 | 27.76 | 23.43 |
| **1 Week** | 20.29 | 22.08 | 25.38 | 23.60 | 20.29 |
| **3 Days** | 20.22 | 20.33 | 21.88 | 20.50 | 20.22 |

- 20 Hidden Layers;
- No restrictions on the maximum number of Training Epochs;
- 2 hour delays in the storage method;
- 1 month and a half of Solar Irradiance data to train the network.

In section 5.3 we state the results of the final simulations of ANN in comparison with the KNN final results.

## 5.2 The Results of the KNN Tests

Applying the KNN Model to this Case Study, before making the final Forecasts, as it was explained in subsection 4.3.1 all the main parameters must be calculated. The first one is the optimal K parameter where we compare the simple KNN algorithm with the KNN method with Ensembling Learning Approach (ELA), with and without the Weighted Distance Function. In this tested we used one month of records to train the method and predict the next hour in each one of the days. The simulations ended after a 24 hour prediction. The results are displayed on Tables 5.6 and 5.7. The simulation time was also recorded and can be observed in Figure 5.3.

With these results, it is possible to conclude that using the Weighted Distance Function improves the prediction in the great majority of the cases compared to the results obtained without a Weighted Distance Function. Nevertheless, it is not possible to choose the best K parameter based on the information

Table 5.6: Performance of the Best cases in the K Parameter and Distance Method Test in the different predictions

| | K Neighbors | RMSE $[W/m^2]$ | MAE $[W/m^2]$ | MAPE | MAPE$_{model}$ |
|---|---|---|---|---|---|
| | | **01-Feb** | | | |
| **Not-Weighted** | 11 | 48.70 | 36.10 | 0.2218 | 0.3127 |
| | ELA | 69.44 | 58.42 | 0.3265 | 0.4605 |
| **Weighted** | 18 | 52.79 | 39.60 | 0.2362 | 0.3331 |
| | ELA | 75.21 | 63.78 | 0.3636 | 0.5127 |
| **Persistence** | | 102.59 | 85.00 | 0.7091 | |
| | | **23-Feb** | | | |
| **Not-Weighted** | 1 | 89.42 | 72.30 | 0.2160 | 0.5325 |
| | ELA | 100.33 | 89.53 | 0.2542 | 0.6267 |
| **Weighted** | 1 | 89.42 | 72.30 | 0.2160 | 0.5325 |
| | ELA | 92.79 | 77.94 | 0.2320 | 0.5719 |
| **Persistence** | | 127.26 | 114.70 | 0.4056 | |
| | | **08-Apr** | | | |
| **Not-Weighted** | 1 | 94.30 | 69.83 | 0.2382 | 0.3772 |
| | ELA | 90.99 | 78.45 | 0.2645 | 0.4188 |
| **Weighted** | ELA | 85.88 | 64.14 | 0.2210 | 0.3499 |
| | 3 | 88.93 | 71.90 | 0.2359 | 0.3734 |
| **Persistence** | | 153.96 | 118.42 | 0.6316 | |

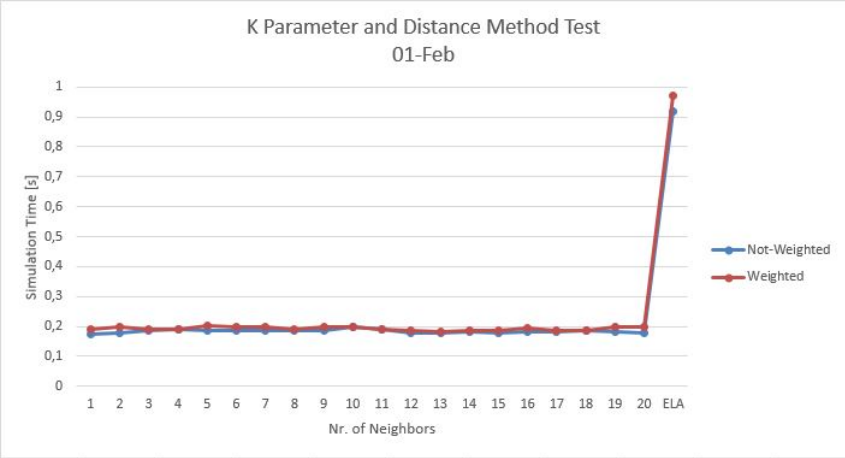
Table 5.7: Cases with best average MAPE$_{model}$ in the K Parameter and Distance Method Test

| **Not-Weighted** | | **Weighted** | |
|---|---|---|---|
| **Number of Neighbors** | **MAPE$_{model}$** | **Number of Neigbors** | **MAPE$_{model}$** |
| 1 | 0.4828 | ELA | 0.4782 |
| ELA | 0.5020 | 1 | 0.4828 |
| 2 | 0.5102 | 2 | 0.4828 |
| 3 | 0.5204 | 4 | 0.4865 |
| 8 | 0.5334 | 3 | 0.5027 |

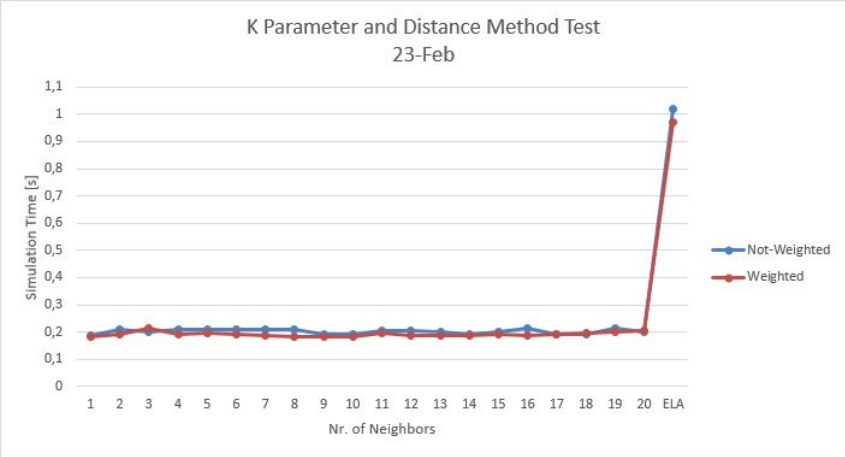
given on table 5.6 because the results are quite different depending on the day that was predicted. As the studies reviewed reported, the best K parameter will always depend on the train set used in the KNN model.

To choose the best case we made the comparison shown on the table 5.7. The case that presents the best MAPE$_{model}$ is the ELA using the Weighted Distance Function. This case is the best case only in one of the three cases (8$^{th}$ of April) but it is the case that presents the best average prediction in all the three days.

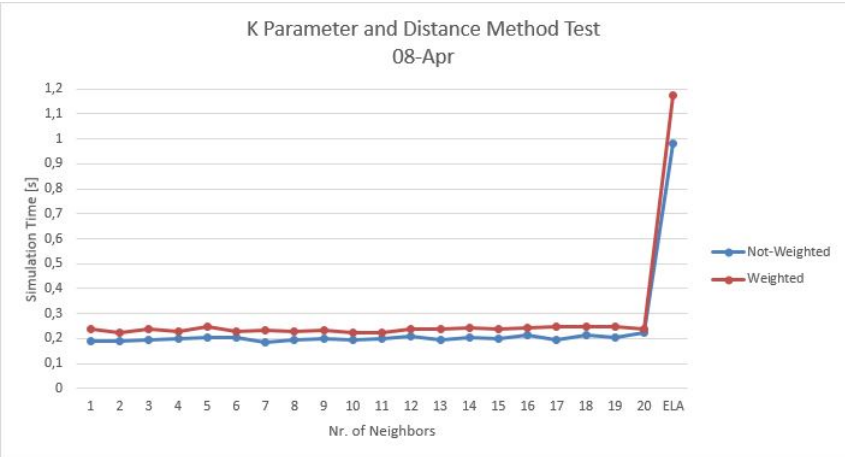
About the Simulation Time analysis of this test we can conclude that all the cases tested run the simulation in a time period very close to each other (about 0.2 seconds). We can verify that for the 1$^{st}$ of February and for the 8$^{th}$ of April, the Not-Weighted cases are in the most part quicker than the Weighted




(a) Simulation Time for forecast of $1^{st}$ of February




(b) Simulation Time for forecast of $23^{rd}$ of February




(c) Simulation Time for forecast of $8^{th}$ of April

Figure 5.3: Simulation Time of KNN K-Neighbors and Distance Method Test for the three forecasts.

cases. The opposite can be observed on the $23^{rd}$ of February where the Weighted cases are in the most cases quicker than the Not-Weighted.

Comparing the ELA with the Normal KNN Algorithm it is evident that the ELA is slower than the Normal

KNN Algorithm. This is due to the fact that ELA predicts the same Solar Irradiance for one hour using several number of nearest neighbors, while the Normal KNN Algorithm predicts the next record using only a fixed value for the K parameter.

For these reasons, while the difference of Simulation Time is not really significant (just 0.8 seconds between ELA and the Normal Algorithm), it was chosen to proceed this work using the ELA and a Weighted Distance Function.

Proceeding this work with the last decision, the Tables 5.8 and 5.9 show the results of the Record Length Test, and the Table 5.10 shows the records of Simulation Time for this test. This test was made under the same conditions as the previous test.

Table 5.8: Results of the Record Length Test for the KNN Model

| | **Record Length** | | | |
|---|---|---|---|---|
| **Length** $[days]$ | **RMSE** $[W/m^2]$ | **MAE** $[W/m^2]$ | **MAPE** | **MAPE**$_{model}$ |
| | | **01-Feb** | | |
| Persis: | 102.59 | 85.00 | 0.7091 | |
| 2 | 67.46 | 53.15 | 0.3275 | 0.4618 |
| 1 | 75.21 | 63.78 | 0.3636 | 0.5127 |
| 3 | 79.41 | 62.57 | 0.3666 | 0.5170 |
| 5 | 94.30 | 76.78 | 0.3998 | 0.5639 |
| 4 | 93.33 | 79.16 | 0.4251 | 0.5995 |
| | | **23-Feb** | | |
| Persis: | 127.26 | 114.70 | 0.4056 | |
| 3 | 36.97 | 34.08 | 0.0934 | 0.2304 |
| 1 | 92.79 | 77.94 | 0.2320 | 0.5719 |
| 2 | 116.22 | 84.95 | 0.2719 | 0.6704 |
| 4 | 167.05 | 132.05 | 0.3221 | 0.7941 |
| 5 | 174.04 | 143.05 | 0.3436 | 0.8471 |
| | | **08-Apr** | | |
| Persis: | 153.96 | 118.42 | 0.6316 | |
| 1 | 85.88 | 64.14 | 0.2210 | 0.3499 |
| 2 | 79.15 | 60.54 | 0.2292 | 0.3629 |
| 5 | 80.60 | 67.25 | 0.3306 | 0.5234 |
| 3 | 143.78 | 101.42 | 0.4579 | 0.7250 |
| 4 | 171.49 | 126.89 | 0.5638 | 0.8927 |

Table 5.9: Comparison Between the MAPE of the KNN Prediction and the Persistence Model of the Record Length Test

| | **Average** |
|---|---|
| Record Length $[days]$ | MAPE$_{model}$ |
| 1 | 0.4782 |
| 3 | 0.4908 |
| 2 | 0.4984 |
| 5 | 0.6448 |
| 4 | 0.7621 |

Table 5.10: Record of Simulation Time in each simulation of the Record Length Test for the KNN Model

| **Records Length** | | | | | |
|---|---|---|---|---|---|
| **01-Feb** | | **23-Feb** | | **08-Apr** | |
| Length $[days]$ | Sim. Time $[s]$ | Length $[days]$ | Sim. Time $[s]$ | Length $[days]$ | Sim. Time $[s]$ |
| 1 | 0.9240 | 1 | 0.9695 | 1 | 1.0217 |
| 2 | 1.6501 | 2 | 1.6798 | 2 | 1.6930 |
| 3 | 2.3696 | 3 | 2.4282 | 3 | 2.3904 |
| 4 | 2.9786 | 4 | 3.2306 | 4 | 2.9811 |
| 5 | 3.6630 | 5 | 3.5611 | 5 | 3.7999 |

The first conclusion that can be drawn by analyzing the table 5.8 is that with bigger records the worst is the prediction. In fact, the case that uses records with 4 days length is the worst in 2 of the three cases, being in some of those cases worse than the Persistence model. Focusing now on the best cases we can verify that the same problem of the last test is present in this one. It is not possible to determine the best case because different day predictions present different best record lengths. Because of that we made the same analysis for the average MAPE$_{model}$ for all three predictions. The results are clear and show that the best case is using records with 24 hours. In fact this case, not being the best in all the three predictions, is the one that presents a better average performance (the best case on the 8$^{th}$ of April and second best case on the remaining days).

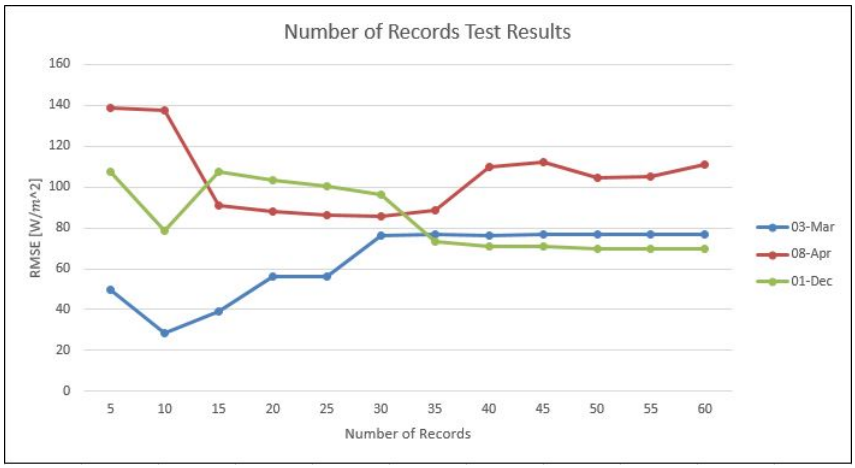
When it comes to the Simulation Time (Table 5.10) it is not a surprise to note that the bigger records the slower the prediction will be. Thus, the case with records 24 hours long is not only the case that has lower error (47.82% of Persistence MAPE) but also the quicker between the five tested cases.

After establishing the best size for the records it is time to evaluate the optimum number of records to predict the next value of Solar Irradiance. The performance results for the Number of Records tests are exposed in Figure 5.4 and 5.5 and in table 5.11.

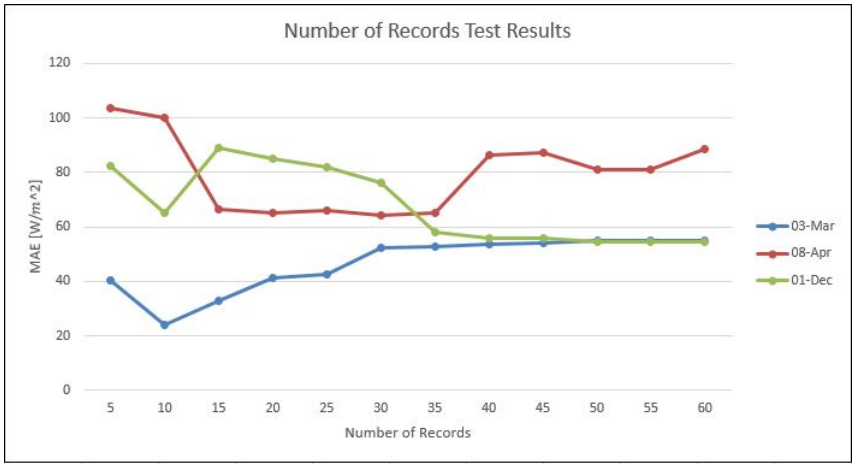
Table 5.11: MAPE Comparison of KNN Model and Persistence Model of the Records Number Test

| **Average** | |
|---|---|
| **Number of Records** $[days]$ | **MAPE**$_{model}$ |
| 35 | 0.4210 |
| 30 | 0.4324 |
| 50 | 0.4588 |
| 60 | 0.4637 |
| 55 | 0.4653 |
| 40 | 0.4660 |
| 45 | 0.4694 |
| 25 | 0.4752 |
| 15 | 0.4776 |
| 20 | 0.4832 |
| 10 | 0.4951 |
| 5 | 0.5118 |

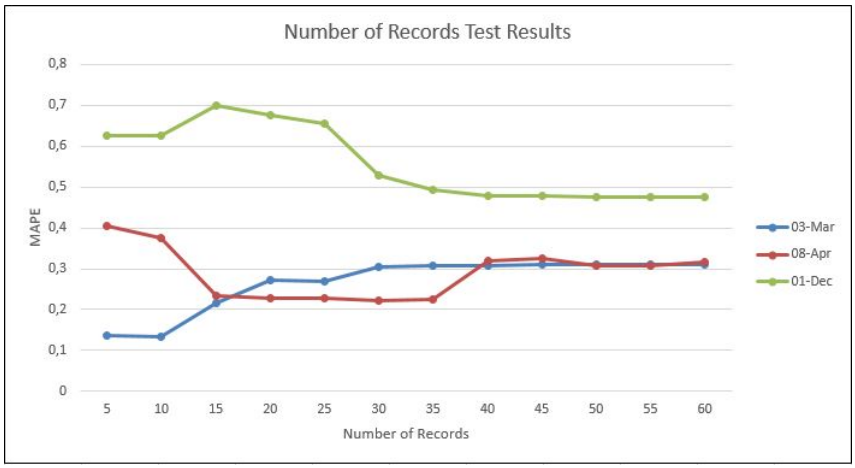
As it was explained in the subsubsection 4.3.1, in this test we predicted different days, because, for




(a) RMSE of the three Forecasts in the Number of Records Test




(b) MAE of the three Forecasts in the Number of Records Test




(c) MAPE of the three Forecasts in the Number of Records Test

Figure 5.4: Results of Number of Records Test for the three forecasts.

example, it is not possible to predict the $1^{st}$ of February with 60 records. For that reason different days with similar daily profiles of Solar Irradiation were chosen.

As in the other tests, the different training sets led to different results. The prediction of the $3^{rd}$ of

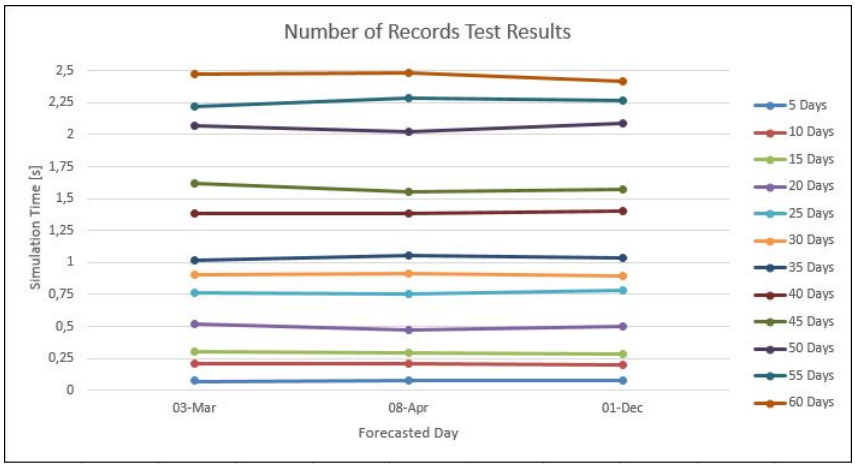



Figure 5.5: Simulation Time for each case tested in Records Number Test for KNN Model.

March favored training matrices with less records, the $8^{th}$ of April favored training matrices with an intermediate number of records and the $1^{st}$ of December favored matrices with higher number of records.

In most cases of this test, the calculated errors are below the Persistence Model Error but to find the best Number of Records of the Training Matrix it was necessary to compare the MAPE of the prediction with the MAPE of the Persistence Model (Table 5.11). Except for the 5 Days case all other cases present predictions below 50% of the Persistence prediction. The best case among all the ones tested was the 35 Days case.

As you would expect, in terms of Simulation Time, the cases with higher amount of records are the cases with higher simulation times. The 35 days case, having an intermediate amount of records, presents a medium simulation time compared with the other cases, which is a very acceptable result.

To sum up, and finishing the KNN Tests, we are able to reunite the best parameters of KNN to predict future Solar Irradiance values:

- the chosen algorithm to train the model and calculate the K-Nearest Neighbors are an ELA with an Weighted Distance Function;

- the records have a length of 24 hours;

- the training matrix has 35 records.

This completes the Parameters Test of the studied models and gathers the best conditions to carry out the comparative simulations of both models. The results of those comparative simulations will be exposed in the next section.

## 5.3 Performance of the Models

After having chosen the best model parameters to perform the best predictions possible, the next step will be to make normal day predictions to compare the performance of both methods. We will carry out two different tests to make this comparison. The first one is intended to analyze the performance of the models in different seasons of the year, and the second one to evaluate the performance of the models in a Clear Sky Day or in a cloudy Day. In order to make this comparison the most accurate possible, the predicted days were the same for both methods and both tests. We remember that in the calculation of the performance errors the night period was not considered.

For the first test we have chosen to predict four specific days of the four different seasons. For training each model, we have chosen the central months of each season (January for Winter, April for Spring, July for Summer and October for Autumn), since these months have the best characteristics of the season. The days chosen to predict were the first days of the last months of the seasons ($1^{st}$ of February for Winter, $1^{st}$ of May for Spring, $1^{st}$ of August for Summer and $1^{st}$ of November for Autumn). In the particular case of the Winter, since both methods need more than 31 days to train the respective models it was not possible to forecast the Solar Irradiance for the $1^{st}$ of February as we said before. Because of that, instead of forecast the intended day it we predicted the $14^{th}$ of February. The results of these comparative tests can be observed in table 5.12 and in the Figures 5.6 to 5.9.

Table 5.12: Performance of Models:seasons; Daily Average

| **Method** | **RMSE** $[W/m^2]$ | **MAE** $[W/m^2]$ | **MAPE** | **MAPE**$_{model}$ | **Sim. Time** $[s]$ |
|---|---|---|---|---|---|
| | | **Winter (14-Feb)** | | | |
| **ANN** | 74.79 | 55.28 | 0.3043 | 0.4791 | 38.60 |
| **KNN** | 57.72 | 43.52 | 0.2409 | 0.3794 | 1.11 |
| **Persistence** | 111.05 | 104.40 | 0.6351 | 1.0000 | |
| | | **Spring (01-May)** | | | |
| **ANN** | 26.32 | 19.44 | 0.0804 | 0.1057 | 42.70 |
| **KNN** | 100.58 | 84.73 | 0.2314 | 0.3042 | 1.20 |
| **Persistence** | 145.35 | 130.36 | 0.7605 | 1.0000 | |
| | | **Summer (01-Aug)** | | | |
| **ANN** | 33.66 | 25.69 | 0.0658 | 0.1427 | 35.45 |
| **KNN** | 31.88 | 22.99 | 0.0415 | 0.0900 | 1.09 |
| **Persistence** | 136.04 | 121.79 | 0.4610 | 1.0000 | |
| | | **Autumn (01-Nov)** | | | |
| **ANN** | 91.02 | 74.33 | 0.4612 | 0.6426 | 38.70 |
| **KNN** | 60.39 | 45.43 | 0.2852 | 0.3974 | 1.10 |
| **Persistence** | 107.58 | 87.80 | 0.7177 | 1.0000 | |

The first observation we can make after the comparative test is that both methods perform a better forecast than the Persistence Method. In some cases it reaches the range of less than 10% of MAPE. The ANN method presents results under the range of 30% of MAPE, excepting the forecast of $1^{st}$ of

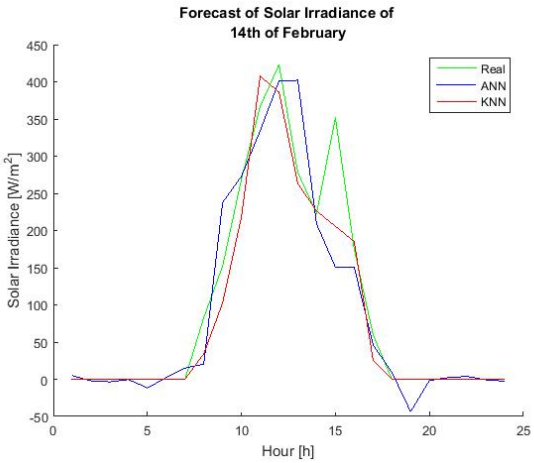



Figure 5.6: Prediction of the 14$^{th}$ of February comparing both methods.

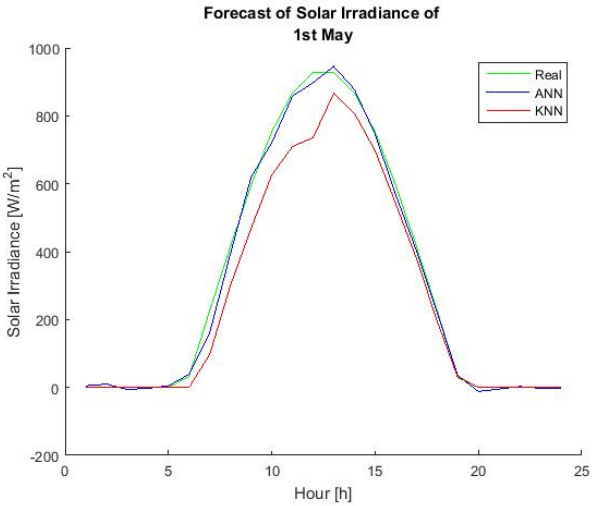



Figure 5.7: Prediction of the 1$^{st}$ of May comparing both methods.

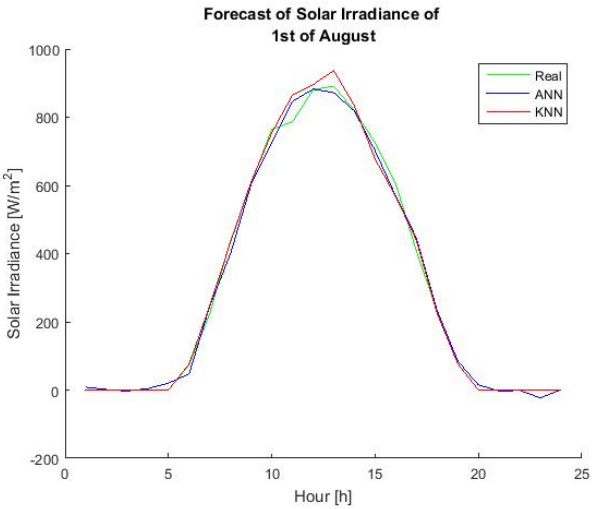



Figure 5.8: Prediction of the 1$^{st}$ of August comparing both methods.

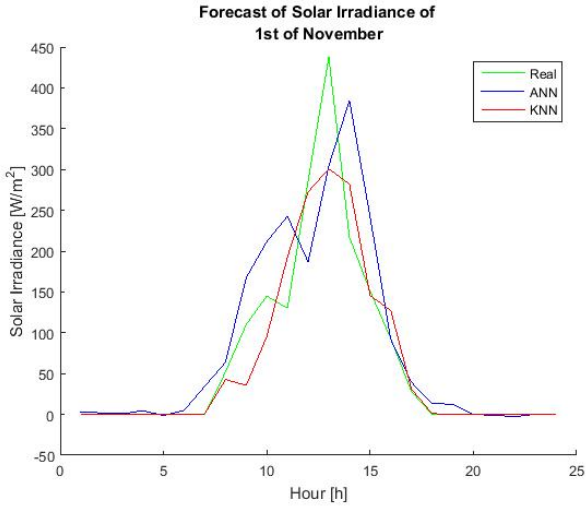



Figure 5.9: Prediction of the $1^{st}$ of November comparing both methods.

November which is the worst prediction with a MAPE of 46.12%. The KNN case shows error values lower than ANN method, except for the prediction of $1^{st}$ of May in which the KNN presents a MAPE of 23.14% while the ANN method presents a MAPE of 8.4%. In general, the period of Spring and Summer presents better results due to the regularity of the Solar Irradiance. Besides this, the forecast of the Solar Irradiance in the seasons of Autumn and Winter is hampered by the irregularity of the Solar Irradiance behavior.

Taking a more detailed analysis of the $14^{th}$ of February forecast we can see that none of the models predict the second solar peak that occurs between the $14^{th}$ and the $16^{th}$ hour. Comparing this case with the Autumn Forecast, that have two peaks as well, we can conclude that maybe this happens due to the quick variation of short duration. The second peak of $14^{th}$ of February only lasts two hours and the models are not well prepared (or trained) for this kind of variations. If we take more attention to the behavior of each model in this period we can observe that, besides they don't reach the real value of Irradiation, both models react to this variation (in the ANN case the forecast for hour 15 and 16 are equal).

In what concerns to the Simulation Time, the KNN method is clearly quicker when compared with the ANN case. The KNN method, like in the parameters tests, performed the predictions in about 1 second while the ANN method needed between 35 and 42 seconds to perform the same predictions.

In the second test fourteen days were randomly chosen, seven Clear Sky Days and seven Cloudy Days. The criteria used to choose these fourteen days was the behavior of the Solar Irradiance. If the Solar Irradiance profile for a certain day was too irregular we classify it as a Cloudy Day. If the profile had a normal behavior increasing the Solar Irradiance from the Sunrise until the Noon and then decreasing until the Sunset without big irregularities then we classified it as a Clear Sky Day. The results of this test can be observed in Figures 5.10 and 5.11 and in Table 5.13. The Figures 5.12 and 5.13 presents the best cases of prediction for both models in each situation. The Figures A.1 to A.3 present in Appendix

A show the worst cases for the same situations.

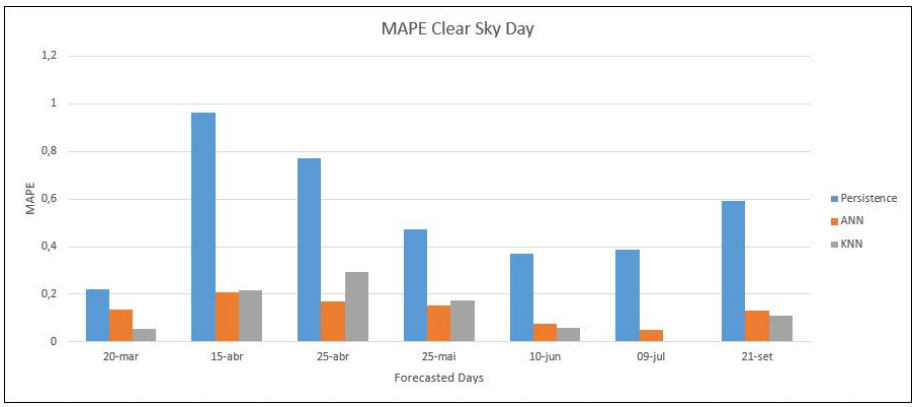



Figure 5.10: MAPE error for the Clear Sky Day Test comparing the studied methods.

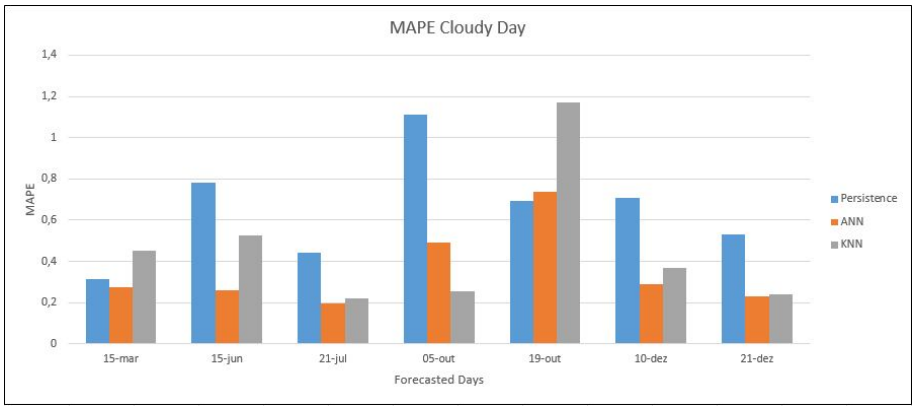



Figure 5.11: MAPE error for the Cloudy Day Test comparing the studied methods.

Table 5.13: Clear and Cloud Day Test 7 Days Average Results

| | **Clear Sky Day** | | | | **Cloudy Day** | | | |
|---|---|---|---|---|---|---|---|---|
| | **RMSE** $[W/m^2]$ | **MAE** $[W/m^2]$ | **MAPE** | **Sim. Time** $[s]$ | **RMSE** $[W/m^2]$ | **MAE** $[W/m^2]$ | **MAPE** | **Sim. Time** $[s]$ |
| **Persistence** | 134.60 | 112.31 | 0.5406 | 0.00 | 101.79 | 86.70 | 0.6537 | 0.00 |
| **ANN** | 51.42 | 37.93 | 0.1319 | 44.79 | 84.67 | 64.63 | 0.3549 | 39.17 |
| **KNN** | 48.77 | 36.60 | 0.1305 | 1.06 | 123.67 | 90.33 | 0.4616 | 1.08 |

The first conclusion to be drawn from the results is that, as the previous test showed us, both methods are in average better than the Persistence Model. In fact, the MAPE for both models was always below 50%. The only exception is the RMSE and the MAE of KNN to the Cloudy Day Tests which is the worst of the three models. In the Clear Sky Day test, as in the test of the Seasons, the KNN model presented in average better results with MAPE of 13.05% (in four of the seven forecasts the KNN presented the best result). In the Cloudy Days case, in contrast to what we saw in the previous tests, the best method in average is the ANN with a MAPE of 35.49% (in five of the seven forecasts the ANN model made a more accurate prediction than the others).

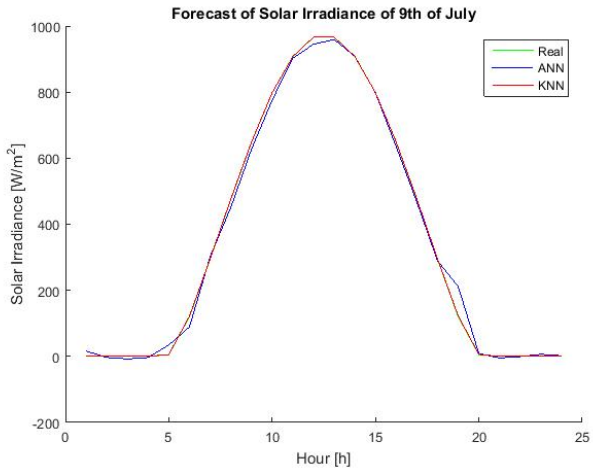



Figure 5.12: Clear Sky Day Forecast: Prediction of the 9$^{th}$ of July comparing both methods. ANN's and KNN's best Clear Sky Forecast.

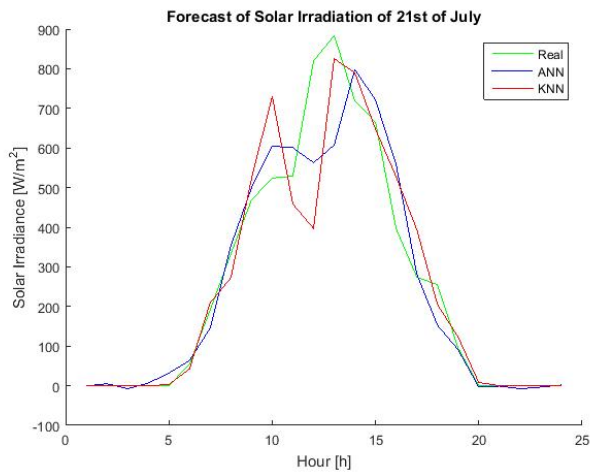



Figure 5.13: Cloudy Day Forecast: Prediction of the 21$^{st}$ of July comparing both methods. ANN's and KNN's best Cloudy Day Forecast.

In this test it was very clear the dependency of the KNN model on the targets associated to past records. As we can see in Figure 5.11 the forecast made by the KNN model reached a MAPE error higher than 100% in the forecast of 19$^{th}$ of October, which is rare for the KNN model. This can be justified with the fact that the training set was not suitable to this case. In fact, if all the last 30 days reached an average Solar Irradiance of 400$W/m^2$ and the day we want to forecast will reach a maximum of 200$W/m^2$ the model can't adjust itself to this variation from the training set. If we analyze the performance of ANN, besides the MAPE of around 70% for the same day, the model could adjust itself more easily than KNN and at least present a result close to the Persistence model. This ANN characteristic was the main responsible for the better result in comparison with the KNN for the Cloudy Day Forecast test.

Regarding the simulation time, the results obtained in these tests were similar to the previous ones. The ANN model presented an average of 44.79$s$ for the Clear Sky Test and 39.17$s$ for the Cloudy Day Test, while the KNN model presented for the same tests an average of 1.06$s$ and 1.08$s$ respectively. This

difference is due to the greater need of computational memory and perform many iterations to predict the next irradiance value. While the ANN perform many iterations and an unknown number of them the KNN perform always the same number of distance calculations in each forecast.

# Chapter 6

# Conclusions

Throughout this work it was possible to learn more about the Artificial Intelligence Techniques, in particular about the ANN and the KNN models and how these models could be applied to the Solar Irradiance Forecast.

In the beginning of the work we could understand the environmental reasons that make the commitment to RES a global need and the work that has already been made by the international governments and the Portuguese government as well. The chapter ends with a summary of the work already been done in the Solar Irradiance Forecast area. Then, it was presented a theoretical view of both models, where was discussed the points that served as inspiration for the creation of these models, the mathematical foundations that support them and the particularities of the models studied in this work.

After this, it was developed two different algorithms using the studied methods ANN and KNN to perform Solar Irradiance Forecast and before the final tests, the models were tuned in order to perform the best forecasts.

After these small optimizations, the models were subjected to two tests. The first one to understand the performance of the models on seasonal days, training with elements of one season of the year to predict another day of the same season. The second test was intended to evaluate the performance of both methods for a typical clear day profile and a typical cloudy day profile.

The results on the performance tests showed us that both methods are better alternatives to the simple Persistence Model. In terms of results, both methods obtained average forecast errors below 30% of MAPE. The only exception was for the Cloudy Day tests where they got errors between 50% and 30%. The KNN showed the best results by surpassing ANN performance in every tests except for the Cloudy Day Test where he obtained an average of 46.16% and the ANN obtained an average of 35.49%. It is worth noting by the positive the great forecasts of the Models in the Spring and Summer forecast (ANN predicted with errors below 10% in both days and KNN also predicted below the same result in

the Summer day) and in the Clear Day test with forecast errors below 15% for both methods.

Relatively to the Simulation Time both methods were consistent with the results obtained in the Parameters test. The KNN were the quickest of the studied models by performing his forecasts in an average time of one second. The ANN performed in average his forecasts between 45 and 35 seconds. This difference can be explained by the different complexity level of the each algorithm and the different memory needs of each model. While KNN only needs to compute the distance between the training points and the input point the ANN needs to repeat the Learning Algorithm several times until reach the prediction for certain hour. Also the operation are different, while in KNN the operations are simply the distance calculation and the average of solar irradiances of the k-Neighbors, the ANN performs more complex operations like derivatives, the propagation of the input through the network or the computation of the Jacobian Matrix.

This work was important also to understand the vulnerabilities of both methods in performing Solar Forecast. In the first test was possible to see that these methods have some difficulties to predict small and quick variations in Solar Irradiance during the day. As we have seen in the previous chapter, when these variations occur the models are slow to react and to correct their predictions. In some cases is possible to see the change in the slope of the predictions. The faster this variation is, the worse it will be for the models to correct. In the second test we could see another handicap for these models, especially for the KNN. These models are dependent on the targets used to train them. If these targets belong to a range of values and the day we want to predict does not even reach irradiances half of these values, then the prediction made for this day will present high errors. This can be well explicit in the KNN forecast for the 19$^{th}$ of October (Figure A.3) where it reach errors of 100% MAPE. This disadvantage is not so explicit in the ANN due to is capability of adapting and learning from the information given. To conclude, if we want a fast model that uses few computational resources regardless of its handicap on depending on the targets used to train, the KNN is the recommended method. If time and resources are not a problem and we want to favor a more adaptive method that presents good results in the accuracy of forecasts, the ANN is the best method.

The main objective of this work was to compare different Artificial Intelligence techniques in order to find the method who performs the best Solar Forecast. For this reason, a next step in future work could be develop and test new optimizations in these two methods. Optimizing the initial choice for the Weights Network in order to reach more quickly the best prediction on the ANN, or develop different training sets for different meteorologic conditions for KNN are good suggestions of future work. In this work we only studied two models, the ANN and the KNN, but as we have seen in 2.3 there are other techniques who can perform a solar irradiation forecast. In addition to other techniques, in the case of ANN other architectures and learning algorithms are an hypothesis to study. Finally, in most of the works it was used solar irradiance values to train and predict the next value, but it is also interesting to study the effect of including in the models the prediction of other meteorologic parameters related with the solar irradiance

(e.g. temperature, cloud percentage) in order to improve the accuracy prediction and to prepare the models for any abrupt variations of solar irradiance.

# Bibliography


[1] R. Castro, *Uma Introdução às Energias Renováveis: Eólica, Fotovoltaica e Mini-Hídrica*, 1st ed. IST Press, 2011.

[2] (2017, February) Facts: Vital Signs - Global Temperature. NASA. [Online]. Available: https://climate.nasa.gov/vital-signs/global-temperature/

[3] (2017, February) Facts: Vital Signs - Sea Level. NASA. [Online]. Available: https://climate.nasa.gov/vital-signs/sea-level/

[4] (2017, January) Kyoto Protocol. United Nations Framework Convention on Climate Change. [Online]. Available: http://unfccc.int/kyoto_protocol/items/2830.php

[5] (2017, April) 2020 climate & energy package. European Comission. [Online]. Available: https://ec.europa.eu/clima/policies/strategies/2020_en

[6] (2017, April) Paris Agreement. European Comission. [Online]. Available: https://ec.europa.eu/clima/policies/international/negotiations/paris_en

[7] G. Carr. (2012, November) Sunny uplands: Alternative energy will no longer be alternative. The Economist. [Online]. Available: http://www.economist.com/news/21566414-alternative-energy-will-no-longer-be-alternative-sunny-uplands

[8] R. E. P. N. for the 21st Century, "Renewables 2016, Global Status Report," Renewable Energy Policy Network for the 21st Century, Tech. Rep., 2016.

[9] I. E. Agency, "Energy Policies of IEA Countries; Portugal; 2016 Review," International Energy Agency, Tech. Rep., 2016.

[10] P. do Conselho de Ministros, "Resolução do Conselho de Ministros nº20/2013," Diário da Repúbica, 1ªSérie - nº70, April 2013.

[11] D. G. de Energia e Geologia, "Renováveis: Estatísticas Rápidas," Ministério da Economia, Tech. Rep. 146, December 2016.

[12] J. Shi, W. J. Lee, Y. Liu, Y. Yang, and P. Wang, "Forecasting Power Output of Photovoltaic Systems Based on Weather Classification and Support Vector Machines," *IEEE Transactions on Industry Applications*, vol. 48, no. 3, pp. 1064–1069, May 2012.

[13] C. Monteiro, T. Santos, L. A. Fernandez-Jimenez, I. J. Ramirez-Rosado, and M. S. Terreros-Olarte, "Short-Term Power Forecasting Model for Photovoltaic Plants Based on Historical Similarity," *Energies*, vol. 6, no. 5, pp. 2624–2643, May 2013.

[14] N. Bigdeli and A. H. Zandieh, "Solar Irradiance Prediction by a New Forecast Engine Composed Wavelet Packet Transform and Adaptive Neuro-Fuzzy Inference System," *IJECCE*, vol. 4, no. 6, pp. 1598–1606, November 2013.

[15] W. S. McCulloch and W. Pitts, "A Logical Calculus of the Ideas Immanent in Nervous Activity," *Bulletin of Mathematical Biology*, vol. 52, no. 1/2, pp. 99–115, 1990.

[16] F. Rosenblatt, "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[17] A. Yona, T. Senjyu, A. Y. Saber, T. Funabashi, H. Sekine, and C. H. Kim, "Application of neural network to 24-hour-ahead generating power forecasting for PV system," in *2008 IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century*, July 2008, pp. 1–6.

[18] C. Paoli, C. Voyant, M. Muselli, and M.-L. Nivet, "Forecasting of preprocessed daily solar radiation time series using neural networks," *Solar Energy*, vol. 84, no. 12, pp. 2146 – 2160, 2010.

[19] P. Mandal, S. T. S. Madhira, A. U. haque, J. Meng, and R. L. Pineda, "Forecasting Power Output of Solar Photovoltaic System Using Wavelet Transform and Artificial Intelligence Techniques," *Procedia Computer Science*, vol. 12, pp. 332 – 337, 2012.

[20] E. Fix and J. L. Hodges, "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties," *International Statistical Review / Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.

[21] E. Patrick and F. Fischer, "A generalized k-nearest neighbor rule," *Information and Control*, vol. 16, no. 2, pp. 128 – 152, 1970.

[22] S. Yakowitz, "Nearest-neighbour methods for time series analysis," *Journal of Time Series Analysis*, vol. 8, no. 2, pp. 235–247, 1987.

[23] A. T. Lora, J. R. Santos, J. R. Santos, J. L. M. Ramos, and A. G. Exposito, *Electricity Market Price Forecasting: Neural Networks versus Weighted-Distance k Nearest Neighbours*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 321–330.

[24] Y. Zhang and J. Wang, "K-nearest neighbors and a kernel density estimator for GEFCom2014 probabilistic wind power forecasting," *International Journal of Forecasting*, vol. 32, no. 3, pp. 1074–1080, 2016.

[25] H. T. Pedro and C. F. Coimbra, "Nearest-neighbor methodology for prediction of intra-hour global horizontal and direct normal irradiances," *Renewable Energy*, vol. 80, pp. 770 – 782, 2015.

[26] M. Hagan, H. Demuth, M. Beale, and O. D. Jesús, *Neural Network Design (2nd Edition)*. Martin Hagan, 2014.

[27] M. T. Hagan and M. B. Menhaj, "Training Feedforward Networks with the Marquardt Algorithm," *IEEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 989–993, November 1994.

[28] J. S. Marques, *Reconhecimento de Padrões: métodos estatísticos e neuronais*, ser. Ensino da Ciência e da Tecnologia, I. Press, Ed. IST Press, 1999, vol. 8.

[29] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, John Wiley & Sons Inc, 2001.

[30] A. B. Hassanat, M. A. Abbadi, G. A. Altarawneh, and A. A. Alhasanat, "Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach," *(IJCSIS) International Journal of Computer Science and Information Security*, vol. 12, no. 8, August 2014.

# Appendix A

# More Results

In this appendix will be presented more test results that show some relevance to this work.

## A.1 Clear Sky Day and Cloudy Day Test

In this section is presented the worst cases of both models for the Clear Sky Day and for the Cloudy Day Tests.

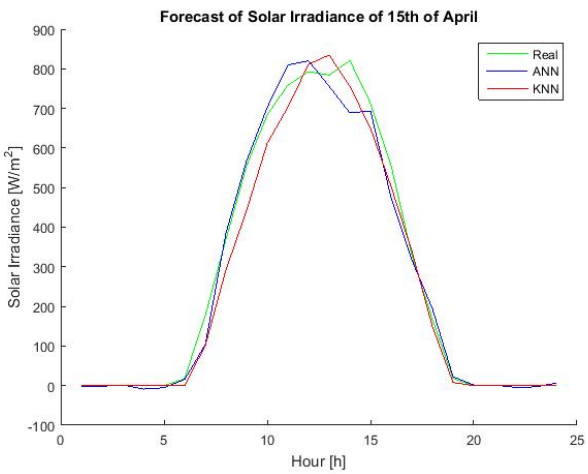



Figure A.1: Clear Sky Day Forecast: Prediction of the $15^{th}$ of April comparing both methods. ANN's worst Clear Sky Forecast.

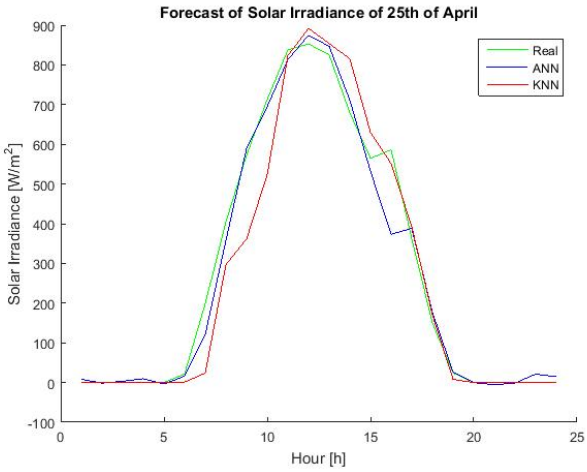



Figure A.2: Clear Sky Day Forecast: Prediction of the $25^{th}$ of April comparing both methods. KNN's worst Clear Sky Forecast.

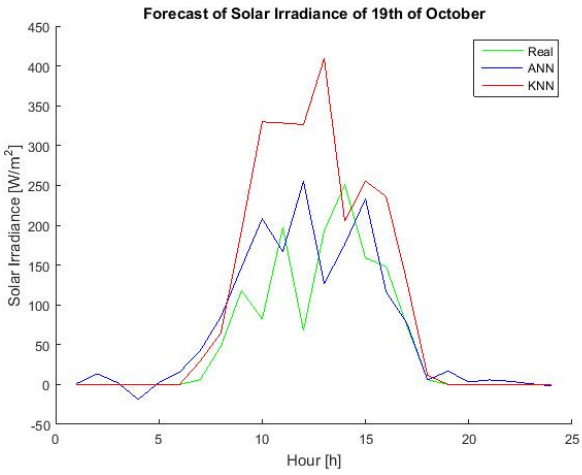



Figure A.3: Cloudy Day Forecast: Prediction of the $19^{th}$ of October comparing both methods. ANN's and KNN's worst Cloudy Day Forecast.

# Appendix B

# Statistics

In this appendix is possible to observe more figures representing important statistics referent to RES and Solar Power.

## B.1 Global Statistics

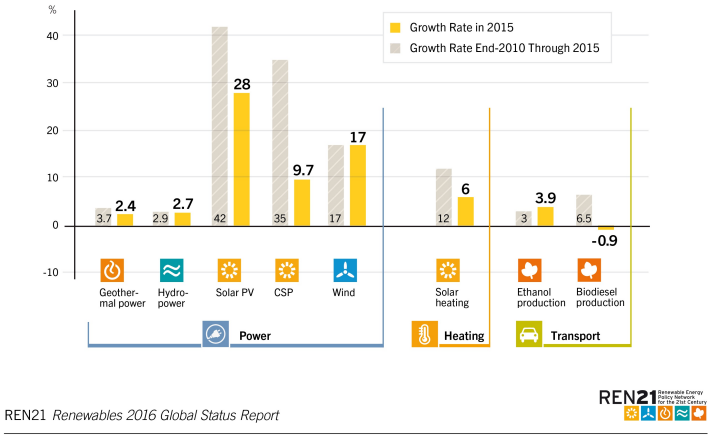



Figure B.1: Graphic representing the Average Annual Growth Rates of Renewable Energy Capacity for all RES and Biofuels Production for the period between 2010 and 2015 and for the year of 2015. Source: REN21, [8].

| | World | China | Brazil | United States | India | Japan | Bang-ladesh | European Union[j] | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Germany | France | Rest of EU |
| | THOUSAND JOBS | | | | | | | | | |
| Solar PV | 2,772 | 1,652 | 4 | 194 | 103 | 377 | 127 | 38 | 21 | 84 |
| Liquid biofuels | 1,678 | 71 | 821[c] | 277[f] | 35 | 3 | | 23 | 35 | 47 |
| Wind power | 1,081 | 507 | 41 | 88 | 48 | 5 | 0.1 | 149 | 20 | 162 |
| Solar heating/ cooling | 939 | 743 | 41[d] | 10 | 75 | 0.7 | | 10 | 6 | 19 |
| Solid biomass[a,g] | 822 | 241 | | 152[e] | 58 | | | 49 | 48 | 214 |
| Biogas | 382 | 209 | | | 85 | | 9 | 48 | 4 | 14 |
| Hydropower (small-scale)[b] | 204 | 100 | 12 | 8 | 12 | | 5 | 12 | 4 | 31 |
| Geothermal energy[a] | 160 | | | 35 | | 2 | | 17 | 31 | 55 |
| CSP | 14 | | | 4 | | | | 0.7 | | 5 |
| **Total** | **8,079[h]** | **3,523** | **918** | **769** | **416** | **388** | **141** | **355[i]** | **170** | **644[k]** |

Note: Figures provided in the table are the result of a comprehensive review of primary (national entities such as ministries, statistical agencies, etc.) and secondary (regional and global studies) data sources and represent an ongoing effort to update and refine available knowledge. Totals may not add up due to rounding.

[a] Power and heat applications (including heat pumps in the case of the EU). [b] Although 10 MW is often used as a threshold, definitions are inconsistent across countries. [c] About 268,400 jobs in sugar cane and 190,000 in ethanol processing in 2014; also includes 200,000 indirect jobs in equipment manufacturing and 162,600 jobs in biodiesel in 2015. [d] Equipment manufacturing and installation jobs. [e] Biomass power direct jobs run to only 15,500. [f] Includes 227,562 jobs for ethanol and 49,486 jobs for biodiesel in 2015. [g] Traditional biomass is not included. [h] The total for 'World' is calculated by adding the individual totals of the technologies, with 3,700 jobs in ocean energy, 11,000 jobs in renewable municipal and industrial waste and 14,000 jobs in others (jobs that cannot be broken down by technology). [i] All EU data are from 2014, and the two major EU countries are represented individually. [j] Includes 8,300 jobs in publicly funded R&D and administration; not broken down by technology. [k] Includes 8,000 jobs in renewable municipal and industrial waste and 3,700 jobs in ocean energy.

Figure B.2: Figure exposing by numbers the amount of Jobs in Renewable Energy Worlwide by Industry. Source: REN21, [8].

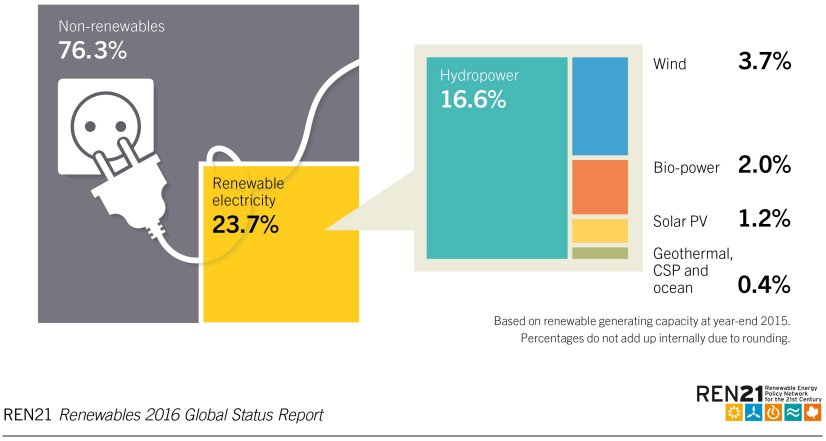



Figure B.3: Graphic representing the estimated Share of Renewable Energy of Global Electricity Production at year-end of 2015. Source: REN21, [8].

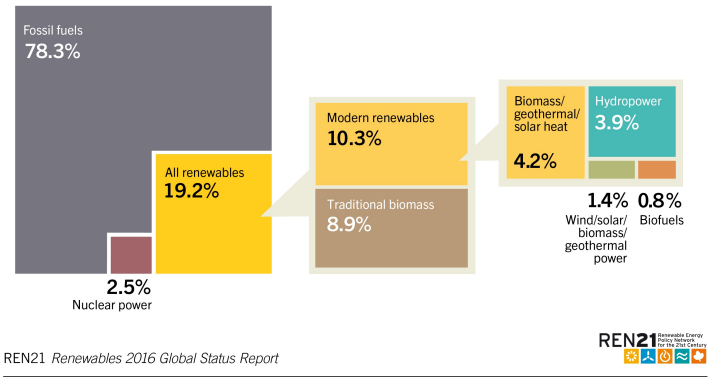



Figure B.4: Graphic representing the estimated Share of Renewable Energy of Global Final Energy Consumption in 2014. Source: REN21, [8].

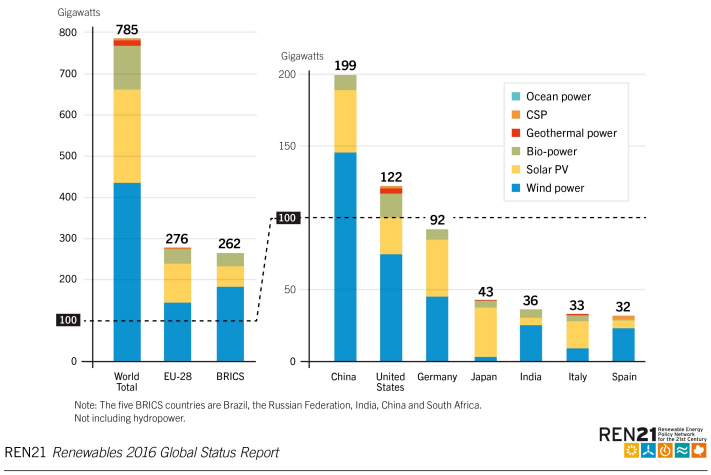



Figure B.5: Graphic representing the Renewable Power Capacities in World biggest groups and in the Top seven Countries at the end of 2015. Source: REN21, [8].

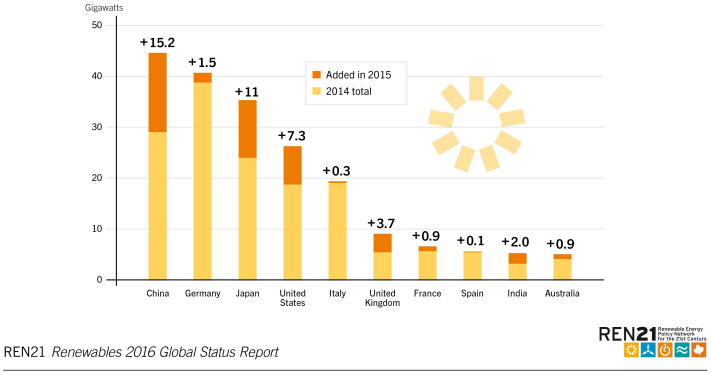



Figure B.6: Graphic representing the Top 10 Countries with the most Solar PV Capacity and their additions in 2015. Source: REN21, [8].

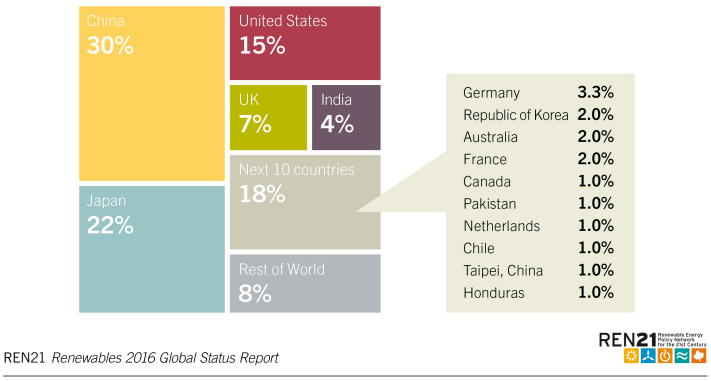



Figure B.7: Graphic representing the Top 15 countries with the biggest Solar PV Global Capacity Addition Shares. Source: REN21, [8].

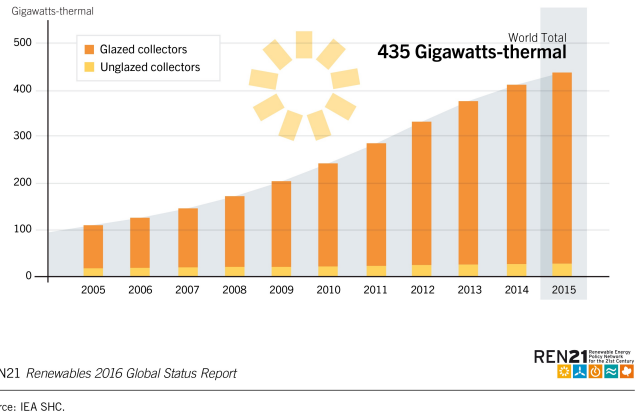



Figure B.8: Graphic representing the world's evolution of Solar Water Heating Collectores Global Capacity between 2005 and 2015 . Source: REN21, [8].

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Investment in renewable power and fuels (not including hydro > 50 MW) | **China** | United States | Japan | United Kingdom | India |
| Investment in renewable power and fuels per unit GDP | **Mauritania** | Honduras | Uruguay | Morocco | Jamaica |
| Geothermal power capacity | **Turkey** | United States | Mexico | Kenya | Germany/Japan |
| Hydropower capacity | **China** | Brazil | Turkey | India | Vietnam |
| Solar PV capacity | **China** | Japan | United States | United Kingdom | India |
| Concentrating solar thermal power (CSP) capacity | **Morocco** | South Africa | United States | – | – |
| Wind power capacity | **China** | United States | Germany | Brazil | India |
| Solar water heating capacity | **China** | Turkey | Brazil | India | United States |
| Biodiesel production | **United States** | Brazil | Germany | Argentina | France |
| Fuel ethanol production | **United States** | Brazil | China | Canada | Thailand |

REN21 *Renewables 2016 Global Status Report*

Figure B.9: Figure Exposing the Top 5 Countries in Investment, Capacity Additions in each RES and Biofuel Production in 2015. Source: REN21, [8].

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **POWER** | | | | | |
| Renewable power (incl. hydro) | **China** | United States | Brazil | Germany | Canada |
| Renewable power (not incl. hydro) | **China** | United States | Germany | Japan | India |
| Renewable power capacity *per capita* (among top 20, not including hydro) | **Denmark** | Germany | Sweden | Spain | Portugal |
| Biopower generation | **United States** | China | Germany | Brazil | Japan |
| Geothermal power capacity | **United States** | Philippines | Indonesia | Mexico | New Zealand |
| Hydropower capacity | **China** | Brazil | United States | Canada | Russia |
| Hydropower generation | **China** | Brazil | Canada | United States | Russia |
| CSP | **Spain** | United States | India | Morocco | South Africa |
| Solar PV capacity | **China** | Germany | Japan | United States | Italy |
| Solar PV capacity *per capita* | **Germany** | Italy | Belgium | Japan | Greece |
| Wind power capacity | **China** | United States | Germany | India | Spain |
| Wind power capacity *per capita* | **Denmark** | Sweden | Germany | Ireland | Spain |
| **HEAT** | | | | | |
| Solar water heating collector capacity | **China** | United States | Germany | Turkey | Brazil |
| Solar water heating collector capacity *per capita* | **Austria** | Cyprus | Israel | Barbados | Greece |
| Geothermal heat capacity | **China** | Turkey | Japan | Iceland | India |
| Geothermal heat capacity *per capita* | **Iceland** | New Zealand | Hungary | Turkey | Japan |

REN21 *Renewables 2016 Global Status Report*

Figure B.10: Figure Exposing the Top 5 Countries in Total Capacity or Generation as of the end of 2015. Source: REN21, [8].