UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ENGENHARIA GEOGRÁFICA, GEOFÍSICA E ENERGIA

Decentralized Solar PV generation forecast based on peer-to-peer approach

Luís Carlos Claudino Teixeira da Silva

Mestrado Integrado em Engenharia da Energia e Ambiente

Dissertação orientada por:
Miguel Centeno Brito (FCUL)
Margarida Pedro (EDP)

2018

Acknowledgment

I wish to acknowledge those who, either by chance or simple misfortune, were involved in my dissertation.

I want to begin by thanking my mentors, Rodrigo Amaro and Professor Miguel Brito for the trust, the support given throughout the project and for helping me to overcome all the doubts that have appeared.

I would also like to thank Eng. Margarida Pedro and Eng. Filipa Reis and the entire team from the EDP Inovação, for the opportunity of this project and for accepting me.

To my family, for being my support. Especially my mother, with her big words, never let me down.

Finally, my great friend, João Costa, for giving me the strength and support along this work. And to all my colleagues and friends who were present for me, like Francisco Pinto and Nuno Pego.

Resumo

Atualmente, cerca de 80% de energia consumida provêm de fontes não renováveis. O aumento da concentração de gases de efeito de estufa e poluentes na atmosfera está fortemente relacionado com a conversão de combustíveis fósseis em eletricidade, calor e a sua utilização em veículos. Este percurso leva a mudanças climáticas que vão perturbar o delicado equilíbrio dos ecossistemas e pode facilmente levar várias espécies de animais e plantas à sua extinção.

A utilização de fontes de energia renováveis como solar ou o vento para a produção de energia elétrica pode fornecer uma solução fazível para a redução de emissão de gases de efeito de estufa, já que são livres de emissões poluentes para a atmosfera, e desta forma salvaguardar futuras gerações. A energia solar, em particular produção fotovoltaica (PV), pode satisfazer as necessidades elétricas de toda a humanidade. Na última década, sistemas PV têm vindo a aumentar o seu potencial, com a integração de células PV em edifícios de habitação e serviços. Em 2015, o mercado PV ultrapassou vários recordes com a sua expansão global, tendo nesta altura uma capacidade total instalada de cerca de 227 GW, dez vezes maior do que em 2009. Em 2016 voltaram-se a bater recordes com um aumento de cerca de 75 GW.

Em Portugal a tecnologia PV tem sido, de todas as tecnologias renováveis, a que mais cresceu, sendo praticamente inexistente em 2006 para uma capacidade instalada de 467 MW no fim de 2016. Cerca de 72 MW estão instalados no sector residencial. Na Europa cerca de 20 GW está instalado no sector residencial, para um total da ordem de 50 GW (no mundo).

A introdução desta nova tecnologia na rede elétrica levanta desafios devido ao seu carácter não controlável, causando vários problemas para os operadores da rede elétrica, como por exemplo problemas de fluxo de potência inversa ou de controlo de voltagem.

As previsões de radiação solar e/ou produção fotovoltaica proporcionam um modo para os operadores de rede controlarem e gerirem os balanços de produção e de consumo de energia de forma a otimizar o processo e não terem de usar as suas reservas de segurança, reduzindo custos.

Com o aumento de produção PV distribuído, a necessidade de prever a sua potência produzida é cada vez maior para o operador do sistema distribuição (DSO). Um dos principais objetivos desta dissertação é estudar e desenvolver um modelo de previsão para a produção PV distribuída na perspetiva do DSO. Isto significa que o modelo de previsão criado deve poder ser aplicado a qualquer sistema sem saber os seus detalhes técnicos; os sistemas analisados vão ser considerados como caixas negras, onde apenas a potência instalada, a distância entre sistemas e o seu histórico de produção é conhecido.

Existe uma grande variedade de técnicas de previsão. Como tal foi feita uma filtragem destas técnicas com o objetivo de ter um modelo simples como base de comparação, um modelo eficiente e um modelo mais complexo, mas com um grande potencial. Após uma análise de bibliografia os modelos escolhidos foram a Persistência (usada como um bom modelo de referência), regressões multivariáveis (simples, com resultados bastante satisfatórios) e o modelo de vetores de suporte (que têm tido um acréscimo de desenvolvimento e resultados bastante interessantes).

A Persistência é o modelo que assume que o que está a acontecer no presente se repetirá no futuro. Embora simples, este modelo pode obter melhores resultados do que modelos numéricos de previsão de tempo (NWP) já que para horizontes curtos consegue traduzir uma melhor representação das nuvens do que os modelos de NWP. Nos casos de horizontes mais alargados

(algumas horas ou dias) este modelo começará a obter erros bastantes grandes, mas serve uma base de comparação apropriada para outros modelos.

O modelo de regressões multivariáveis (ARX) é um modelo de regressão linear que incorpora mais do que uma variável independente. Na maioria dos casos as relações principais costumam ser lineares ou, caso não sejam, podem ser assumidas de uma forma satisfatória, e desta forma aplicar este modelo e obter resultados interessantes. As grandes vantagens deste modelo é a sua fácil utilização e implementação, mas pode levantar dificuldades quando temos poucos dados ao fazer um ajuste demasiado fino aos dados de treino e perder a sua capacidade de generalizar para variáveis ainda não conhecidas.

Um modelo mais complexo que também usa regressões é o modelo de vetores de suporte. Este modelo foi inicialmente um modelo orientado para problemas de classificação linear, para separar duas classes através de uma linha reta num plano em outra dimensão, chamado de Hyperplane. Este modelo foi mais tarde usado para problemas não lineares onde foi adaptado para problemas de previsões através de regressões lineares por Vapnik, 1995, passando a chamar-se regressão de vetores de suporte (SVR).

O conceito geral deste modelo é transformar os dados fornecidos para uma outra dimensão onde seja possível fazer uma regressão linear, e depois devolver esta função. Esta transformação para outra dimensão é bastante facilitada graças a facilidade de incorporar o uso de funções *kernel* já existentes. O único problema é a escolha dos vários *kernel* que se torna um problema não trivial. Para este trabalho escolheu-se o kernel RBF, o mais usado por outros estudos semelhantes. Os parâmetros principais deste modelo são o *épsilon* e o *cost*. Estes parâmetros vão ser coeficientes que vão criar um intervalo onde o que esteja dentro deste será ignorado pelo modelo, olhando este apenas para os que restaram; estes dados que não são ignorados são chamados de vetores de suporte.

Outra forma deste modelo é o nu SVR, que em vez de nos deixar escolher o *épsilon*, permite-nos escolher a percentagem de vetores de suporte a usar e desta forma ele automaticamente calcula o *épsilon* a usar. A vantagem desta abordagem é na poupança de tempo computacional a escolher o melhor *épsilon*, mas com a contrapartida de perder algum controlo sobre o erro do modelo, já que não é possível escolher o *épsilon*. A grande vantagem do modelo SVR é que de certa forma é imune ao excesso de ajuste dos dados para o treinar já que os parâmetros *épsilon* e *cost* evitam isto e consegue obter desempenhos satisfatórios com o uso de poucos dados.

Para a previsão de radiação solar ou produção fotovoltaica é importante considerar o índice de céu limpo. O índice de céu limpo (K) relaciona a condição de um instante totalmente limpo (sem nebulosidade) com a condição real existente. Este índice varia entre K=1, que é um instante sem nebulosidade, e K=0, que representa um instante sem qualquer luz (noite, ou eclipse). O K pode ser usado para isolar a interferência causada pelas nuvens e eliminar os impactos da posição dos módulos ou da variabilidade sazonal. A remoção destes impactos permite a utilização da informação de vários sistemas diferentes de uma forma equivalente. Neste trabalho o K é calculado através da abordagem de Lonij, pela sua simplicidade e a necessidade de apenas do histórico de produção do(s) sistema(s) selecionado(s).

Nesta dissertação os dados usados para desenvolver e testar os modelos de previsão referem-se a Inglaterra, região de East Midlands, para um ano completo (01/07/2015 até o dia 30/06/2016), com um passo de 30 min. Foi realizada uma análise preliminar a estes dados, eliminando as estações com demasiados valores em falta; no final são usados dados de 57 sistemas. Além da previsão da geração individual de cada sistema, foi também considerado testar a previsão para a geração agregada da região em estudo. Todos os modelos foram desenvolvidos em R.

Um dos primeiros problemas identificados foi o tempo necessário para a aplicação do modelo SVR. Os dados foram transformados para um passo de 1h de forma a reduzir em metade o número de dados usados. Os parâmetros para o SVR foram reduzidos a cinco hipóteses por parâmetro (épsilon, cost e gamma). Os números de estações previstas do modelo foram reduzidos de 57 a 12 (as 57 estações foram usadas como variáveis pelos modelos para a previsão) e o nu SVR foi utilizado para verificar se a redução no tempo era significativa e se haveria perda de desempenho. Os horizontes estudados foram de 1h até 6h.

O modelo da Persistência foi o primeiro a ser testado. Cada modelo terá inicialmente apenas informações do tempo presente (n = 1), sendo o primeiro caso de estudo, depois adicionando 2 pontos passados (n = 3) e, finalmente, adicionando 4 pontos passados (n = 5).

Em todos os casos de estudo, tanto o ARX quanto o SVR superaram o modelo de Persistência, exceto no quarto caso (n=5), o ARX neste caso, especialmente para horizontes maiores, têm o pior resultado. Os resultados em geral demonstraram que o uso de informações das estações próximas ajuda a precisão da previsão.

Concentrando no caso em que se usa apenas dados do presente (n = 1), a previsão do ARX e SVR, mostra um desempenho muito semelhante, com ligeira vantagem para SVR em horizontes maiores. Um aspeto sistemático na previsão para o caso regional é ter um desempenho menor do que o caso individual para horizontes inferiores, mas quase igual para horizontes maiores. Como a referência de comparação de ambos os modelos (ARX e SVR) é a Persistência e esta melhora no caso regional, devido a soma dos vários sistemas suavizar o comportamento da produção PV, os modelos aparentam ter pior performances para horizontes baixos, o que não se verifica no erro médio quadrático normalizado (nRMSE).

Quando n = 3 e n = 5, o SVR supera o ARX, sobretudo para horizontes maiores, mas com pior desempenho do que o caso de usar apenas informação presente (n = 1). Isso significa que adicionar informação passada apenas teve impacto negativo nos modelos. Isto pode ser uma adversidade de usar dados com um passo de 1 hora e não o passo original de 30 minutos, já que desta forma, para horizonte grandes (4h para cima) o modelo ARX pode perder a sua capacidade de generalizar tendo um ajuste demasiado grande ao treino, devido a redução de alvos na previsão. As nuvens podem ser eventos muito rápidos, portanto, para um melhor resultado, dever-se-ia usar dados de resolução mais alta, já que o passo de 1 hora é uma resolução muito baixa. A principal ideia que pode ser retirada destes casos é que o modelo SVR é um modelo mais robusto do que o ARX e pode lidar melhor com o uso de mais variáveis, não sofrendo de excesso de ajustamento, que parece acontecer no modelo ARX.

O modelo SVR exige uma computação muito mais exigente do que o ARX, sendo o principal problema o processo de otimização dos parâmetros. A alternativa utilizada nu SVR, reduziu consideravelmente esse processo sem perda de desempenho, uma alternativa ótima para o caso em questão, mas se realmente for necessário controlar a quantidade de erro no modelo e ir para o melhor desempenho possível, o épsilon SVR é o modelo escolhido. Estes modelos deveriam ser testados com dados de resolução mais elevada, já que deve melhorar ambos os modelos, especialmente o SVR, uma vez que parece que pode superar o problema de excesso de ajustamento melhor, mas tal não foi possível tendo em consideração o computador usado para esta dissertação.

Palavras-chaves: Regressões lineares multivariáveis (ARX), regressão de vetores de suporte (SVR), persistência, potência solar, índice de céu limpo, previsões de minutos/horárias, operador do sistema de distribuição (DSO), R

Abstract

This work studies solar power forecasting based on multivariable linear regressions (ARX) and support vector regressions (SVR) for a set of spatially distributed photovoltaic systems, and their aggregate. Models consider data-driven through a clear sky index from multiple neighbor systems available. The method is applied for very short-forecasting from the perspective of the distributed system operator (DSO). Forecast performance is assessed by comparison with the performance of the Persistence model, by evaluating forecast root mean square error (RMSE). Results for a case study with 57 PV systems in Sheffield, UK, show that in general, the SVR model presented better performances than ARX, especially for longer horizons. It is also shown that the SVR model can handle well overfitting problems. On the other hand, the model requires large computation power and time. The addition of neighbor's information has a positive result in the forecasting performance for all models.

Keywords: Forecasting, persistence, multivariable linear regression (ARX), support vector regression (SVR), solar power, very short-forecasting, distributed system operator (DSO), clear sky index, R

Table of contents

Figures and Table Index

Symbology

PV	Photovoltaic
DSO	Distribution System Operator
TSO	Transmission System Operator
DG	Distributed Generation
NWP	Numerical Weather Prediction
NWS	National Weather Service
GHI	Global Horizontal Irradiance
SAM	System Advisor Model
IDE	Integrated Development Environment
CRAN	Comprehensive R Archive Network
PPF	Past Predicts Future model
NN	Neural Network
AR	Auto Regressive
ARMA	Auto-Regressive Moving Average
ARIMA	Auto-Regressive Integrated Moving Average
ARX	Multiple Linear Regression
SVM	Support Vector Machine
SVR	Support Vector Regression
WSVM	Weighted Support Vector Machine
MSE	Mean Square Error
RMSE	Root Mean Square Error
nRMSE	Normalized Root Mean Square Error
MAE	Mean Absolute Error
NA	Missing Values
W	Watt
kW	Kilowatt
MW	Megawatt
GW	Gigawatt
kWh	Kilowatt hour

kVAh	Kilo-volt-ampere hour
kWp	Kilowatt peak

1. Introduction

This chapter makes a brief introduction to the photovoltaic topic and the motivation for the use of forecasting techniques by system operators (mainly focused on the distributed system operator) for photovoltaic production. The objectives and organization of this work are also presented.

1.1. Context

In presents days, we get approximately 80% of the energy we consume from non-renewable energy sources, e.g. fossil fuels [1]. The increase in emissions of greenhouse gases and pollutants is correlated with the conversion of fossil fuels to electricity, heat, and transport. This leads to global warming that disturbs the delicate ecosystems and could easily relegate many species to extinction [2]. To prevent a global warming calamity, the global temperature should be kept to or around 1.5 ºC above pre-industrial temperatures. For this target, the global greenhouse gas emissions should be reduced worldwide to approximately 80% from their 1990 levels, by 2050 [3].

Utilizing renewable sources energy such as wind and solar to generate electricity provides a feasible contribution for the greenhouse gas reduction challenge as these are emission-free sources of energy that can be used to generate electricity and at the same time protect our environment for future generations.

Solar energy, in particular photovoltaic (PV), can fulfill all the electricity needs of humankind [4]. In the last decade, PV solar energy has started realizing its huge potential, as the amount of installed PV power is rapidly increased, also with the integration of solar cells into the roofs and facades of buildings [5].

After a limited development in 2014, the market restarted its fast growth, almost everywhere, with all regions of the world contributing to PV development for the first time. In 2015, the PV market broke several records and continued its global expansion, with a 25% growth at 50 GW. The total installed capacity at the end of 2015 globally amounted to at least 227 GW, ten times higher than in 2009. In 2016, the worldwide PV market showed again a great potential, breaking again several records and continuing with its global expansion, with a 50% increase [4].

Overall, these developments raised the annual global PV market to at least 75 GW, with a positive outcome in all regions of the world. Photovoltaic has now reached 1 GW of regional penetration on all continents, much more on the leading ones (Asia, Europe, and North America). The total global installed capacity as passed the 300 GW mark, as shown in Figure 1.1 [6].

The increase in the capacity installed yearly has been coupled with a strong decrease in the components price. Since 2006, the PV system price has shown a reduction of more than 50%, e.g. the standard final price in 2006 was around 5500-6000 €/kW for a residential system, whereas in 2017 the standard final price was approximal 2400-2700 €/kW [5].

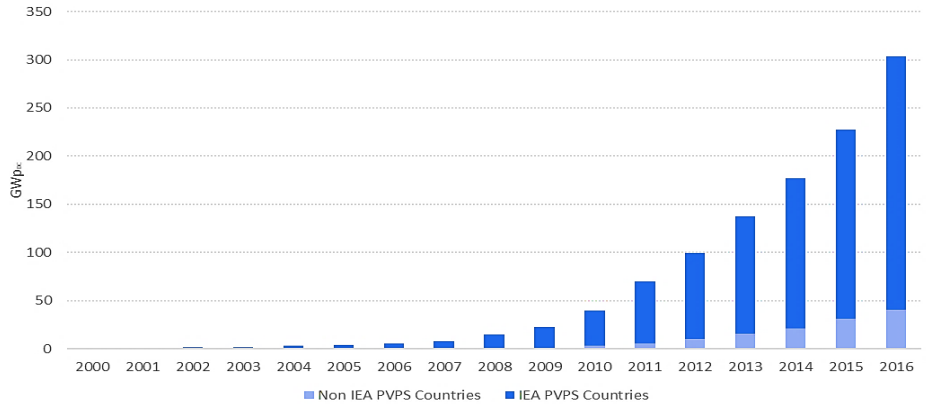

Figure 1.1: PV Installations evolution along the years [2000-2016] [6]

It was thought that integrating PV systems into electrical grids would not be a difficult task; however, when the penetration level of PV systems started to increase, utilities began to face new non-traditional challenges mainly due to the discontinuous nature of solar energy [7]. Photovoltaic array output is highly dependent on environmental conditions such as illumination intensity and temperature. For example, the presence of clouds or high wind velocity can reduce the power output of solar cells and the presence of dust on the surface of the solar panels deteriorates their performances [8]. Thus, the power system must deal with not only uncontrollable demand but also uncontrollable generation [9].

Photovoltaic systems can impose negatives impacts on the electrical grid, dependent on the size as well as the location of the PV system. They may be classified based on their ratings into three different categories: i) Small systems rated as 10 kW and less, ii) intermediate PV systems rated from 10 kW to 500 kW and iii) large PV systems rated above 500 kW. The first two categories are usually managed by the distribution system operator (DSO) and the latter by the transmission system operator (TSO) [10].

Some of the main negative impacts of PV systems on the DSO side are as follows:

Reverse Power Flow

In the distribution system, the power flow is usually unidirectional from the Medium Voltage system to the Low Voltage system. However, at a high penetration level of PV systems, there are moments when the net production is higher than the net demand, especially at noon. As a result, the direction of power flow is reversed, and power flows from the Low Voltage side to the Medium Voltage side. This reverse power flow results in overloading of the distribution feeders and excessive power losses. The reverse flow of power has also been reported to affect the operation of automatic voltage regulators installed along distribution feeders as the settings of such devices need to be changed to accommodate the shift in load center [11].

Utilities can mitigate this adversity by setting a limitation to the PV power to be exported or simply blocking the export since the PV power is typically used for local loads. If a frequent accurate forecast of PV power could be done these limitations could be controlled relatively well [12].

Power Losses

Distributed Generation (DG) systems, in general, reduce system losses as they bring generation closer to the load. This assumption is true until reverse power flow starts to occur.

Miller and Ye showed that distribution systems losses reach a minimum value at a penetration level of approximately 5% but losses increases as the penetration level increases [13]. So, a key challenge is managing reverse power flow from happening.

Voltage Control Difficulty

In a power system with embedded generation, voltage control becomes a difficult task due to the existence of more than one supply point, in this case, will appear situations of overvoltage and under voltage [14].

As an example, the PV output is high when it is sunny, but when clouds appear PV output power can drop very quickly resulting in an under-voltage condition. This problem can be mitigated using weather prediction to reduce the expected amount of output power when clouds are expected. This requires the use of real-time weather measurements and weather forecast algorithms [15].

Increased Reactive Power

Photovoltaics system inverters normally operate at unity power factor for two reasons. The first reason is that standards do not allow PV system inverter to operate in voltage regulation mode. The second reason is that owners of small residential PV systems normally are paid for their kWh yield, not their kVAh production. Thus, they prefer to operate their inverters at unity power factor to maximize the active power generated and accordingly, their return. As a result, the active power requirements of existing loads are partially met by PV systems, reducing the active power supply from the utility. However, the reactive power requirements are still the same and must be supplied completely by the utility. A high rate of the reactive power supply is not preferred by the utilities because in this case distribution transformers will operate at very low power factor and the efficiency of transformers decreases as their operating power factor decreases. Hence overall losses in distribution transformers will increase reducing the overall system efficiency [16].

Islanding

It is necessary to detect when the system operates in an island mode and to disconnect it from the grid as soon as possible. The island can occur when a part of the grid is electrically isolated from the power system but the part with island is energized by distributed generators. The islanding detection is important for many reasons - a possibility to damage customer equipment's and distributed generator, hazard for line-workers, islanding may interfere with restoration of normal services for neighboring customers [17].

For a quality and reliable distribution of power produced by PV systems the ability of precise forecast the energy produced is of great importance and has been identified as one of the key challenges for massive integration [18].

1.2. Motivation

Solar forecasting provides a way for system operators to predict and balance energy generation and consumption. Assuming the system operator has a mix of generating assets at their disposal, reliable solar forecasting allows optimization of dispatch their controllable units. Thus, a proper PV forecast would be able to lower the number of units in hot standby and, consequently, reduce the operating costs.

In this perspective, an inaccurate solar forecast means a need to make up for unpredicted imbalance with shorter-term sources of power. These short-term sources tend to be costlier on a per unit basis, which also means that the extent of total inaccuracy is important. For instance, the total cost to make up a 10% error on a 10 MW and 100 MW plant will, of course, be very different. This cost can then be passed through from the system operator to the market participants.

These types of penalties have existed with controllable generation to assure reliable power delivery to the grid operator but are now becoming an option for renewable plant operators that see economic value in using forecasts to schedule power. Accurate forecasting is required to take on that type of additional risk. So, an accurate forecast is not only beneficial for system operators since it reduces costs and uncertainties, but also for PV plant managers, as they avoid possible penalties that are incurred due to deviations between forecasted and produced energy [19].

From all renewable sources explored in Portugal, PV technology saw the largest relative increase in installed capacity, going from practically inexistent in 2006 to 467 MW in the end of 2016, is approximately 72 MW in the residential sector [20]. This residential growth tendency can be seen in Europe and in the rest of the world, and the tendency is to continue to go up Figure 1.2.

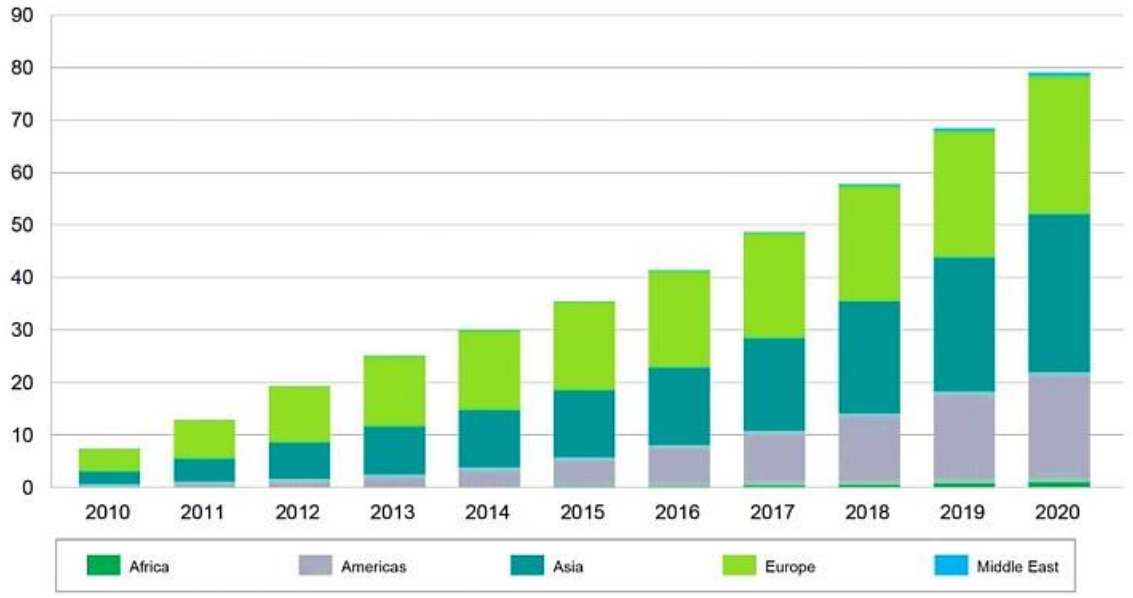

Figure 1.2: Global Cumulative Residential PV Installations (GW) Source: IHS Markit predict 2017 [21]

This residential growth can be attributed mainly due to the reduction in the price of PV systems components, with a focus on batteries, and incentive politics that are done in each country. Even so, the prices for the majority of the population is too high [22].

Given the increasing number of installed small PV systems, new challenges regarding the grid integration of PV generated electricity is coming up. Amongst those challenges is the issue of ramps and peaks of the injected PV power into low voltage grids, as talked in the context topic.

In this new challenge resides the main motivation for the DSO side, being the main objective of this perspective maintain the quality of energy (tension and safety of distribution) and managing the electric network with efficiency, to reduce losses. If the DSO can't guarantee the quality of energy distributed, they need to compensate the customers and the overutilization of the electrical network assets will impact its lifespan. The other barrier is how efficiently the network can be handled since if it is badly managed it will impact directly on the operational cost. Forecasting the

PV power production would help mitigate this problem and create new solutions for load control [23].

Either way, the trustworthy forecasting of the expected PV power production is crucial for the integration of high shares into our energy system

1.3. Objectives

With the increase of PV distributed generation, the necessity to forecast their power output is growing necessity for the DSO operations. So, the main objective of this dissertation is to study and develop a forecasting model for solar PV distributed generation with the perspective of the DSO, responsible for operating, ensuring the maintenance of and, if necessary, developing the distribution system in each area.

The various systems will be analyzed as a black box, with the purpose that the forecasting model can be applied to any system without knowing their technical details. The information available will be the installed power, the location and the historical generation of each system, Figure 1.3 represents an example of this condition. In this example we can see that the systems closer to the clouds are producing less than the system further away, in this manner if the historical data for the generation of each system and their distance could be used in the forecasting model, it could interpret the movement of the clouds, improving the forecast accuracy.

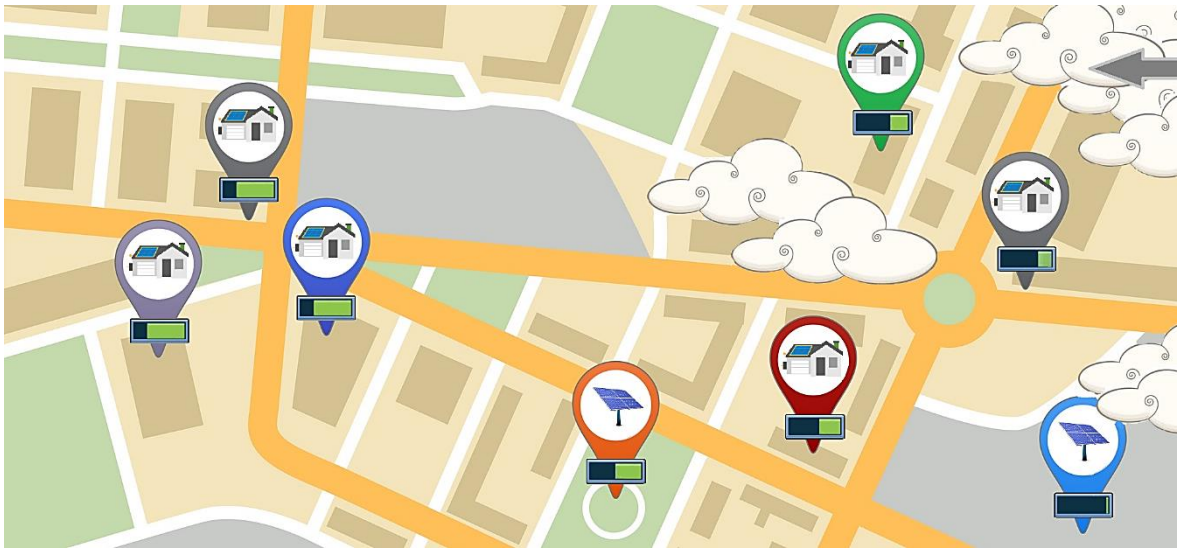

Figure 1.3: Example of various residential PV systems with their location and generation know. The green bar indicates the PV electric production of each system and the wind direction is indicated by the grey arrow at the top right.

Nowadays there is a great variety of forecasting techniques so developing all of them would have been too time and computationally demanding, so with the interest of time and work management three techniques were chosen, Persistence (as a base model for comparison), Multivariable Regression (as a simple and efficient model) and Support Vector Regression (more advanced model with the potential of having superior performances).

1.4. Organization of the dissertation

This chapter introduced the impacts and challenges of large-scale integration of photovoltaics in the grid, the motivation of the work and defining its objectives.

In chapter two the different models used for this work are presented. It is described how forecast methods work in general and a theoretical explanation of each model used, their process, variables, advantages, and disadvantages.

Chapter three is devoted to the methodology used in this work. Knowing that the models had already been used and studied, the objective is not to focus directly on the model but rather how the models react to different variables with an index in forecasting decentralized PV power with a peer-to-peer information.

In chapter four the results and their interpretations are presented, for each figure will be pointed out the analyzes, the test carried out and discussion of the results obtained.

The last chapter (chapter 5) presents a summary and the main conclusions from this work.

2. Solar forecasting

This chapter presents the state of the art of solar forecasting techniques used in this dissertation, persistence, multivariable and support vector regressions, together with a brief explanation of models that use PV power as an index for PV forecast.

2.1. Overview

Accurate power prediction can enhance the stability and security level and lead to a more economical operating decision for the power system. Due to meteorological uncertainty, PV energy is difficult to predict. Weather variables such as temperature, global solar irradiation, sunshine duration, wind speed, relative humidity, cloudiness/sky cover, precipitation and dew point are used as inputs for solar power forecasting models. Solar irradiation varies with time, season, geographical location and meteorological conditions [24]. Since the output power of a solar panel at a fixed temperature is closely linear dependent of the global irradiance [25], predicting solar irradiance is not expected to be very different from predicting PV power, at least in a first approach.

The choice of a model depends on the forecasted horizon, tools, data available, required resolution, accuracy, and purpose. In general, forecasting may be distinguished according to the horizon time scale, but there is no consensus in the literature as to what the thresholds should be. One possible classification [26] is as follows:

- Very short-term generally involves horizons from a few seconds up to a few hours ahead. These services are important for grid operators to guarantee grid stability.
- Short term from a few days to a few months ahead. They are useful in daily operations of utility companies and valuable for electric market operators [27]. Forecasting the power output of a PV pant for the next few hours or days is necessary for the optimal integration in the electric network.
- Long-term with lead times measured in months, quarters or even years. The long-term forecast is necessary for strategic planning. Their information is essential for capacity expansion, capital investment decisions, revenue analysis and corporate budgeting [28], [29].

2.2. Forecasting Models

Forecasting methods can be broadly characterized as physical or statistical [30]. Physical models are based on mathematical equations which describe physical state and dynamic motion of the atmosphere. This approach uses different weather forecasts such as global horizontal irradiance, ambient temperature, relative humidity, wind speed and PV system characteristics such as system location, orientation and historical data or manufacturer specification as inputs of solar and PV models which perform forecast of irradiation in an array plane and back of module temperature [31], [32].

In the statistical approach historical data of PV power and various inputs, such as a ground station or satellite data, numerical weather prediction (NWP) outputs and PV system data are used. A historical dataset is chosen to train a model and to determine unknown model parameters. Models output a forecast of PV power at a given time based on past inputs proceeded by analysis, Figure 2.1.

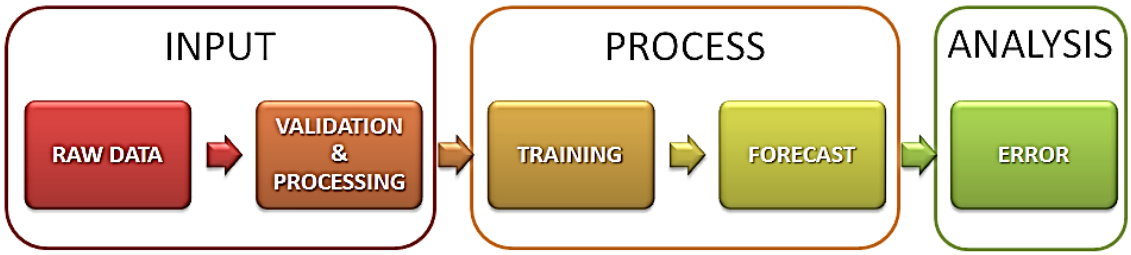

Figure 2.1: Forecast Process for Statistical Approach [33]

A few key components make up solar forecasting tools. First, there is the weather model. As mentioned above, solar generation is variable by nature. Cloud cover causes this variability by obstructing sunlight from hitting the solar panels. If one can predict the weather with a great amount of certainty, one is already one step ahead of improving our forecast.

The second factor in a solar forecast is the model used to convert the weather into system power output. The solar industry uses these "PV simulation" models to predict the performance of a PV system under environmental conditions like irradiance, wind speed, temperature and relative humidity. PV simulation models may also incorporate important PV systems behaviors such as tracking, which predicts the orientation of the PV panels mounted on single- or dual-axis tracking hardware.

Some of the main statistical analysis methods used in power generation forecasting are multiple linear regressions [34], neural networks (NN) [35], support vector machines (SVM) [36], autoregressive moving average (ARMA) [37], autoregressive integrated moving average (ARIMA) [38] for non-stationary time-series.

2.2.1. Persistence

Some simple forecast techniques can serve as benchmarks that can be used to evaluate forecast improvements. One very simple and commonly used reference is the Persistence method. The Persistence model as shown results that can outperform the NWP models, the reason for this is that it can inherit a better representation of the temperature effects on the panels that are not present as an input to the GHI (Global Horizontal Irradiance) to the power output a day-head model [39]. The error in the Persistence increases considerably as the hour-ahead increases, as such the Persistence model is only suitable for very short term forecasting [40].

This method main idea sets that the conditions at the time of the forecast will not change, "things stay the same", for example projecting past values of PV production into the future [4].

2.2.2. Multivariable Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is known as the independent variables (X), and the other is often called the response variable (Y), being B_0 the bias and B the model coefficients, equation (2.1).

$$Y = B_0 + B.X \tag{2.1}$$

The overall idea of regression is to examine two things: (a) does a set of predictor variables do a good job in predicting an outcome variable? (b) Is the model using the predictors accounting for the variability in the changes in the dependent variable? [41].

These regressions are commonly used because they are the simplest and easiest non-trivial relationships to work with, in the majority of the case the key relationships between our variables are often or at least approximately linear over the range of values that are of interest to us and even if they are not, we can often transform the variables in such a way as to linearize the relationships [42].

The multivariable regression is very similar to a linear regression, the difference is in the number of variables we give the model, where we can have multiples explanatory variables, equation (2.2).

$$Y = B_0 + B_1.X_1 + B_2.X_2 + \cdots + B_n.X_n \tag{2.2}$$

Multivariable linear regression analysis has three major uses. First, it can be used to identify the strength of the effect that the independent variables have on the dependent variable. Second, it can be used to forecast effects or impacts of changes, in other words, helps us to understand how much the dependent variable will change when we change the independent variables. Third, this model analysis predicts trends and future values that can be used to get point estimates.

When selecting the model, an important consideration is the model fit. Adding independent variables to multiple linear regression models will always increase the amount of explained variance in the dependent variable (typically expressed as R^2). Therefore, adding too many independent variables without any theoretical justification may result in an over-fit model [41].

When the spatial information is an important variable, auto-regressive (AR) regression can be proven very useful. Overall the auto-regressive model specifies that the output variable will depend linearly on its own previous value and a stochastic term (an imperfectly predictable term). The main advantage of the AR is that the model explicitly accounts for autocorrelation in the error allowing for a valid inference to be made, while normal linear regression assumes uncorrelated errors, with autocorrelated data, this assumption fails and so inference drawn from the regression model will be invalid (standard error, etc)[43].

2.2.3. Support Vector Regression

Support vector machines were, initially, a learning algorithm orientated for linear problems of classification, where they would separate two linear classes in a hyperplane. Later it was proposed for non-linear problems, mapping of points in one space and projecting it for a bigger dimension space, where they are divided according to the class to which they belong, with the intention of maximizing the separation margin. After this, the notion of Support Vector Regression (SVR) was introduced, a generalization of SVM for regression problems (Vapnik, 1995) [36], one of the main advantage of this method is that it can capture the nonlinearity in a dataset.

For a brief explanation of the SVR equations, let's assume that we are given training data $\{(x_1, y_1), \dots, (x_l, y_l)\} \subset \aleph \times \Re$, where $\aleph$ represents the space of the inputs patterns – for instance $\Re^d$. In *ε-SVR* the objective has been to find a function *f(x)* that has at most ε deviation from the obtained targets y_i for all the training data and at the same time as flat as possible. For the case of a linear function *f* has been described in the form as

$$f(x) = \langle \omega, x \rangle + b \quad with \quad \omega \in \aleph, b \in \Re \tag{2.3}$$

where $\langle ... \rangle$ stand for the dot product in $\aleph$. Flatness in (2.3) means small ω. For this, it is required to minimize its Euclidean norm i.e. $\|\omega\|^2$. Regularly this can be written as a convex optimization problem by requiring

$$minimize\ \frac{1}{2}\|\omega\|^2 \tag{2.4}$$

$$subject\ to \begin{cases} y_i - \langle \omega, x_i \rangle - b \leq \varepsilon \\ \langle \omega, x_i \rangle + b - y_i \leq \varepsilon \end{cases} \tag{2.5}$$

The above convex optimization problem is achievable in cases where *f* exists and approximates all pairs (x_i, y_i) with ε precision. Introducing slack variables ξ_i, ξ_i^* (slack variable is a variable that is added to an inequality constraint to transform it into an equality) to cope with the constraints of the optimization problem, (2.4) and (2.5), the formulation becomes

$$minimize\ \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{2.6}$$

$$subject\ to \begin{cases} y_i - \langle \omega, x_i \rangle - b = \varepsilon + \xi_i \\ \langle \omega, x_i \rangle + b - y_i = \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \tag{2.7}$$

The constant C > 0 determines the trade-off between the flatness of *f* and the amount up to which deviations larger than ε are tolerated. ε-insensitive loss function, $|\xi|_\varepsilon$ has been described by

$$|\xi|_\varepsilon = \begin{cases} 0 & \text{if } |\xi| < \varepsilon \\ |\xi| - \varepsilon & otherwise \end{cases} \tag{2.8}$$

The dual formulation provides the key for extending SVM to nonlinear functions. The standard dualization method utilizing Lagrange multipliers has been described as follows:

$$\begin{aligned} L = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) - \sum_{i=1}^{l}\alpha_i(\varepsilon + \xi_i - y_i + \langle \omega, x_i \rangle + b) \\ - \sum_{i=1}^{l}\alpha_i^*\left(\varepsilon + \xi_i^* + y_i - \langle \omega, x_i \rangle - b\right) - \sum_{i=1}^{l}(\eta_i\xi_i + \eta_i^*\xi_i^*) \end{aligned} \tag{2.9}$$

The dual variables in equation (2.9) must satisfy positivity constraints i.e. $\alpha_i,\ \alpha_i^*, \eta_i, \eta_i^* \geq 0$. It follows from saddle point condition that the partial derivatives of L with respect to the primal variables (ω, b, ξ_i, ξ_i^*) should vanish for optimization.

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0 \tag{2.10}$$

$$\frac{\partial L}{\partial \omega} = \omega - \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)x_i = 0 \tag{2.11}$$

$$\frac{\partial L}{\partial \xi_i^*} = C - \alpha_i^* - \eta_i^* = 0 \tag{2.12}$$

Substituting equation (2.10)(2.11)(2.12) into (2.9) yields the dual optimization problem.

$$Maximize \left\{ -\frac{1}{2}\sum_{i,j=1}(\alpha_i - \alpha_i^*)\left(\alpha_j - \alpha_j^*\right)\langle x_i, x_j\rangle - \varepsilon\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) + \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\right\} \tag{2.13}$$

$$Subject\ to \quad \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \quad and\ \alpha_i, \alpha_i^* \in [0, C] \tag{2.14}$$

Dual variables η_i, η_i^* trough condition (2.13) have been eliminated for deriving (2.14). As consequence (2.11) can be rewritten as follow:

$$\omega = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)x_i \tag{2.15}$$

and therefore from Eq. (2.3)

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\langle x_i, x\rangle + b \tag{2.16}$$

This is named support vector expansion i.e. ω can be completely described as a linear combination of the training patterns x_i. Notice that ω does not have to be computed explicitly, even for evaluating $f(x)$.

Computation of b is done by exploiting Karush-Kuhn-Tucker (KKT) conditions [36] which states that at the optimal solution the product between dual variables and constraints must vanish. In this model, means that:

$$\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) = 0 \tag{2.17}$$
$$\alpha_i^*(\varepsilon + \xi_i^* - y_i + \langle w, x_i\rangle + b) = 0 \tag{2.18}$$

and

$$(C - \alpha_i)\xi_i = 0 \tag{2.19}$$
$$(C - \alpha_i^*)\xi_i^* = 0 \tag{2.20}$$

Following this reasoning, only samples (x_i, y_i) with corresponding $\alpha_i^* = C$ lie outside the ε-insensitive tube around f and $\alpha_i\alpha_i^* = 0$, i.e. there can never be a set of dual variables $\alpha_i\alpha_i^*$ which are both simultaneously nonzero as this would require nonzero slacks in both directions. Finally, for $\alpha_i^* \in (0, C), \xi_i^* = 0$ and moreover the second factor in (2.17) and (2.18) need to vanish, hence b can be computed as follows:

$$b = y_i - \langle \omega, x_i\rangle - \varepsilon \quad for\ \alpha_i \in (0, C) \tag{2.21}$$
$$b = y_i - \langle \omega, x_i^*\rangle - \varepsilon \quad for\ \alpha_i^* \in (0, C) \tag{2.22}$$

From (2.17) and (2.18), it follows that only for $|f(x_i) - y_i| \geq \varepsilon$ the Lagrange multipliers may be nonzero, or in other words, for all samples inside the ε-tube, the α_i, α_i^* vanish for $|f(x_i) - y_i| < \varepsilon$ the second factor in (2.17) and (2.18) is nonzero, hence α_i, α_i^* has to be zero such that the KKT conditions are satisfied. Therefore, a sparse expansion of ω exists in terms of x_i (note: not all x_i are needed to describe). The examples that come with non-vanishing coefficients are called Support Vectors (SV).

One important step is turning the SVR algorithm to nonlinear problems and this can be done simply by pre-processing the training patterns x_i by a map $\phi: X \rightarrow \Im$, into some feature space $\Im$ and then applying the standard SVR algorithm. Then the expansion in (2.15) and (2.16) becomes:

$$\omega = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)\phi(x_i) \tag{2.23}$$

and therefore

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)k(x_i, x) + b \tag{2.24}$$

The difference with the linear case is that ω is no longer explicitly given. In the nonlinear setting, the optimization problem corresponds to finding the flattest function in feature space, not in input space. The standard SVR to solve the approximation problem is as follows:

$$f(x) = \sum_{i=1}^{N} (\alpha_i^* - \alpha_i)k(x_i, x) + b \tag{2.25}$$

the kernel function $k(x_i, x)$ has been defined as way to compute the dot products of two vectors x_i and x in some (possibly very high dimensional) feature space.

$$k(x_i, x) = \varphi(x_i)\varphi(x) \tag{2.26}$$

So instead of mapping our data via ϕ and computing the inner product, we can do it in one operation, leaving the mapping completely implicit. This "trick" is called the kernel trick. This process can be implemented so smoothly because of the optimization from the dual formulation and when testing a new example, we only need to sum over the support vectors which is much faster than summing over the entire training-set.

Unfortunately, choosing the 'correct' kernel is a nontrivial task, and may depend on the specific task at hand. No matter which kernel you choose, you will need to tune the kernel parameters to get good performance from your classifier. Popular parameter-tuning techniques include K-Fold Cross Validation [44]. The most used kernels are the polynomial, radial basis functions (RBF) and Saturating (sigmoid-like), Figure 2.2.

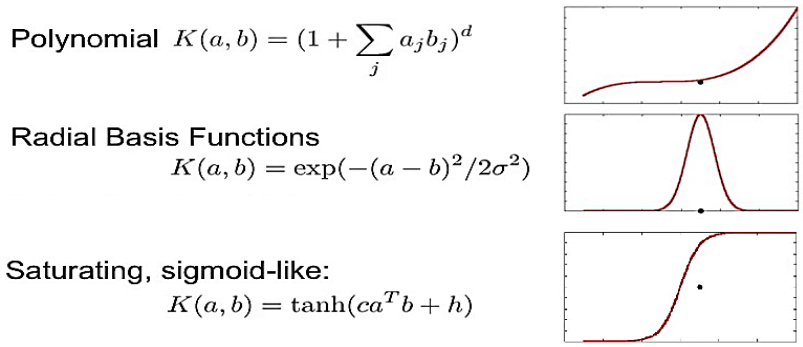

Figure 2.2: Common Kernels [45]

The Lagrange multipliers α_i^* and α_i of (2.25) have been obtained by minimizing the following regularized risk function

$$R_{reg}|f| = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l} L_e(y) \tag{2.27}$$

where the term $\|\omega\|^2$ has been characterized as the model complexity, C as a constant determining the trade-off and the ε-insensitive loss function, $L_e(y)$ given by

$$L_e(y) = \begin{cases} 0, & \text{for } |f(x) - y| < \varepsilon \\ |f(x) - y| - \varepsilon, & otherwise \end{cases} \tag{2.28}$$

In classical support vector regression (ε-SVR), the proper value for the parameter ε is difficult to determine beforehand. Fortunately, this problem is partially resolved in the algorithm, ν support vector regression (ν-SVR), in which ε itself is a variable in the optimization process and is controlled by another new parameter ν ∈ [0,1]. ν is the upper bound on the fraction of error points or the lower bound on the fraction of points inside the ε-insensitive tube, allowing to determine the proportion of the number of support vectors, we desire to keep in our solution, with respect to the total number of samples in the dataset. Thus, a good ε can be automatically found by choosing ν, which adjusts the accuracy level to the data at hand. This makes ν a more convenient parameter than the one used in ε--SVR [46].

However, in ε-SVR you have no control over how many data vectors from the dataset become support vectors, it could be a few, it could be many. Nonetheless, you will have total control of how much error you will allow your model to have, and anything beyond the specified ε will be penalized in proportion to C, which is the regularization parameter.

2.2.4. Clear-sky Index

Solar radiation is always greater in an area that extends perpendicularly to the sunbeams, than in a horizontal area with the same dimensions. Because of earth's rotation and this axis of rotation is tilted, the azimuth and the solar height change throughout the day and the year, the angle of incidence of the solar radiation constantly varies in the areas with the potential to the use of solar energy. In this manner obstacles to the sunlight can have different impacts according to the time of the year, as an example the radiation from a clear-sky mid-day in winter can have the same solar radiation value as a cloudy morning in the summer [47]. As such some of the major challenging factors for PV forecasting are the orientation/inclination of the system since different positions result in different solar profiles and the possible presence of obstacles which provoke shades, originating drastic power variations.

One way to approach this problem would be normalizing radiation measurements to their clearness index, firstly done by J. N. Black, [48]. In this work, the ratio between the daily radiation measurements and they theoretical counterparts in a perfectly transparent atmosphere, called "clearness index", was used to develop regression equations for forecasting daily radiation from sunshine hour observations.

The usefulness of this index was reinforced some years after by B. Y. H. Liu and R. C. Jordan [49], for estimating the performance of tilted flat-plate solar collectors. The methods used before this study needed a detailed record of the radiation and temperature data of the locality of the collector. The fact that a large volume of meteorological radiation and temperature data must be analyzed made the prediction of collector performance an extremely tedious and time-consuming task. The work done in this study made this process easier. They used generalized φ-curve (based in various radiation plots) with the correlation of clearness index, by means of which the performance of a

collector of any angle or tilt at any locality can be predicted when the following two parameters are known: the monthly-average daily total radiation on a horizontal surface and the monthly-average day-time ambient temperature.

A further technique of the clearness index concept has become possible since the development of proficient clear-sky radiation modeling (e.g. Golnas 2011, [50]). By modeling the clear-sky radiation arriving at the surface of the Earth, the denominator of the clearness index (previously the extra-terrestrial component of irradiance) can be replaced with this clear-sky estimate, thereby changing it to the well-known "clear-sky index" which most modern methods now utilize, equation (2.29). This index is a dimensionless number between 0 and 1, has a high value under clear, sunny conditions, and a low value under cloudy conditions.

$$K_{index} = \frac{Measured}{Clear\ Sky\ estimate} \tag{2.29}$$

This model, as handy it can be to calculate the irradiation the system receives, the transition from irradiance to power is not trivial since it implies knowing other variables as, module temperature, inverter efficiency, model efficiency, etc. If these variables are known, models like the one proposed by N. Engerer and F.P.Mills, [51], that uses a horizontal clear-sky radiation model, transposes and then converts to PV power with physical equations, can be a great choice.

A rather different approach is that of Lonij [52] or Bacher [53], that uses a statistic way, in their index, instead of applying to global irradiance it is applied on solar power. In this manner, the application of the clear-sky index to PV power was simplified, and both cases presented interesting forecasting results.

An example where the importance of estimating the clear-sky index is illustrated in Figure 2.3. It is shown that this index ignores changes in the day of the year or system positioning and can also isolate the cloud-induced variability.

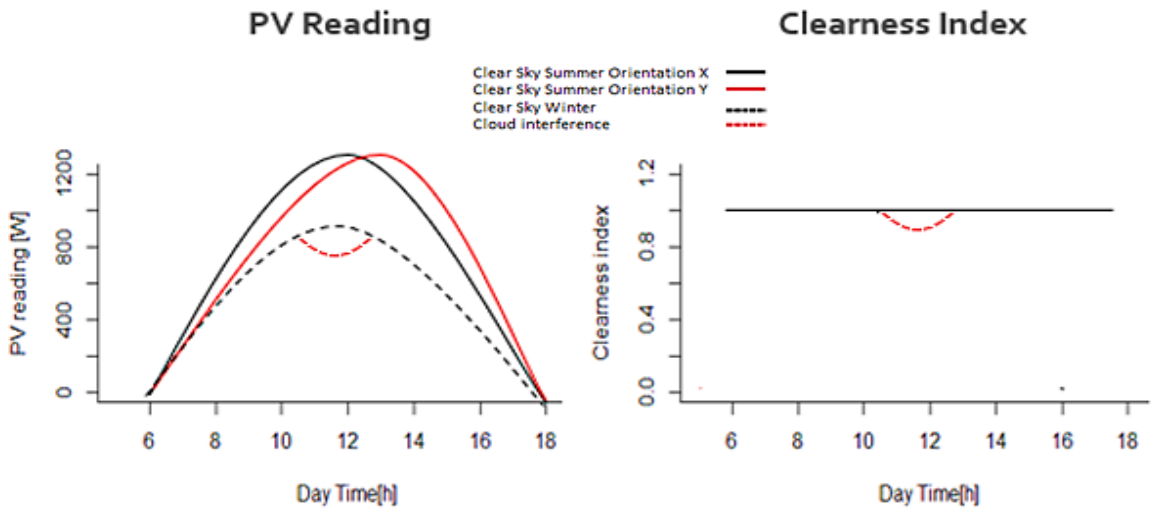

Figure 2.3: On the left the representation for three different system and on the right the K index result

2.3. Studies Overview

Achieving high accuracy forecasts at each time scale imposes specific requirements to the applicable data sources, solar irradiation models, and forecasting techniques converting available data into quality solar power forecasts [29].

Kostylev and Pavlovski, [29], also Huang, [40], describe forecasting methods depending on the tools and information available to forecasters. This includes PV system data, proprietary models such as total sky image based or satellite-based cloud cover and irradiation forecasts and publicly available results NWP models.

Kostylev and Pavlovski [29] states that forecast accuracy strongly depends on the climatic conditions at the forecast site since the cloud regime strongly defines the success of the forecast performance. The example given is from Central European stations, the relative root mean square error (RMSE) ranged from 40% to 60% (% of mean observed), while for sunnier Spanish stations relative RMSE values were in the range of 20% to 35%.

Huang, [40], results demonstrate that the ARMA model is suitable for short and medium term forecasting and the Persistence have good results in short previsions. The ARMA procedure includes obtaining the historical solar radiation data from SolarAnywhere and using System Advisor Model (SAM) for the historical solar generation data. The maximum mean square error (MSE) for the ARMA model was 0.028 kW (in March), showing improvements over of the persistence of 44.38% (in January), this was done for forecasting of 1-hour ahead horizon for a laboratory-level micro-grid scenario.

Elke [30] describe and evaluate the approach of irradiance forecasting, which is the basis for PV power prediction. They present an approach to derive weather specific prediction intervals for irradiance forecasts. First, site-specific hourly forecasts are derived from the low-resolution forecasts of the European Centre for Medium-Range Weather Forecasts. In a second step, the forecast of the global horizontal irradiance must be converted to the module plane with a tilted irradiance model. Finally, the power output forecast is obtained by applying a PV simulation model to the forecasted irradiance. The model returns the alternating current power feed to the grid as a function of the incoming irradiance and the ambient temperature. They showed that the forecast errors are smaller than 5% of the nominal power in more than 80% of all situations and smaller than 10% of the nominal power in more than 90% of all situations.

Sharma [54] study developed prediction models using historical NWS (National Weather Service) forecast data and correlated them with generation data from solar panels. Their analysis quantifies how each forecast parameter affects the other and the solar intensity. For solar energy harvesting, they found that sky cover, relative humidity, and precipitation is highly correlated with each other and with solar intensity, while temperature, dew point, and wind speed are only partially correlated with each other and with solar intensity. They studied how solar intensity varies with individual forecast parameters and how these forecast parameters are related to each other. They also showed how a day of a year affects solar intensity. For this, they used linear least squares and SVM using multiple kernel functions. Their results concluded that the RMSE for SVM-RBF with four dimensions is 128 W/m^2, while the RMSE for cloudy and past predicts future model (PPF) is 175 and 261 W/m^2, respectively. Thus, SVM-RBF with four dimensions is 27% more accurate than the simple cloudy model and 51% more accurate than the PPF model.

Oudjana and co-authors [55] analyze the relationship between meteorological factors and a power supply. They study a design of PV power generation forecasting systems for one week ahead using

weather databases including the global irradiance, and temperature using a data acquisition system. They present three forecasting models using simple regression and neural network methods based on artificial intelligence.

The study done by Xu and co-workers [56] adopts a weighted SVM to forecast the short-term PV power. They selected the five most similar days to the day to be forecasted as the training samples. The weights of the samples for the weighted Supported Vector Machine (WSVM) are designed based on similarities and the time point. The solar irradiation and the temperature were considered as the two main factors in the process of PV power forecasting. Therefore, a history of data including solar irradiation, temperature, and the output power was used to find the similar days. Increasing the number of similar days decreased the prediction accuracy other than improving it. In this work, they also used artificial NN to compare methods. The result showed that WSVM had a maximum of relative error of 12.23% versus the NN maximum relative error of 32.18% (the relative error here is the ratio of Mean Absolute Error (MAE) to the real value), being WSVM with a much smaller error than NN.

Another important feature for forecasting models is the information given to the training set, Kariniotakis and co-authors [57] showed that the accuracy increases with the training period size. Since the model can detect more patterns and relationships between the information provided and thus the model can have a better calibration. The near-optimal performance was reached with a training period of about 20 days. This procedure was done for the Random Forest model to a 200 kWp plant located in the south-east of France. The results had values around 10% RMSE (been normalized by the solar plant nominal power)

Pelland and co-authors [4] report on a general analysis of the state of the art of forecasting techniques, and states that the best day-ahead solar and PV forecasts combine NWP forecasts with postprocessing of these forecasts to improve them or to generate forecasts that are not included in the direct model outputs of the NWP, such as PV forecasts. Key post-processing approaches are a spatial-temporal interpolation, smoothing, and model output statistics. Also, it describes an example of non-parametric statistical methods (Random Forest and SVM) application to data obtained from a real, grid-connected PV power station located in France, demonstrating that these models managed to properly predict sunny days up to a satisfactory accuracy degree, whereas cloudy or unstable days pose more difficulties to be forecast. It his highlighted that the overall performance obtained is twice better than the typical performance for the case of wind farms at flat terrain.

With the increasingly centralized and decentralized PV generation, the more important is to know how to handle this variable source of energy production. Obtaining good forecasting performances is of a great help in handling this challenge since it can provide the information needed to control and maintain a trustworthy use of this energy in our grid as the economic implications it may have. The results obtained in various studies are promising but still not enough for a mass implementation, for this, the research done to forecasting topic is increasing and developing each year with slow but steady progress.

This work will be focused on developing some of the forecasting techniques already used, mainly SVR, using the clear sky index model in Lonij ([52]) with the extra condition that the system technical information is unknown but using the neighbors as a source of information for the model (a peer-to-peer approach). As such the next chapter will explain the process used to develop this approach.

3. Methods

Three different forecasting models were developed: Persistence, Multivariable Regression (ARX) and SVR. The models were implemented in R (an open-source programming language) with the clear sky model. This chapter presents the details of the different methods, describes how they were implemented and how the quality of the outputs was assessed.

3.1. Data

The data used for testing and construction of the R script was originated from a region in the United Kingdom. These data contained the information from one year of PV production with a step of 30 min between each reading, beginning in the day 01/07/2015 to the end of the day 30/06/2016.

A brief analysis revealed that there are stations with a lot of missing values (NA) in certain months and the PV production was the accumulated value over the year. Stations with NAs were removed, giving the option to work with a full complete year of readings and the accumulated value was uncompressed getting the production at given time.

After these filters, the data to work with was that of 57 stations, represented in Figure 3.1. These stations were used as input variables in each model for the forecast of each individual stations.

Figure 3.1: Location and Distance Representation of the Data Systems

In the cases of machine learning algorithms, a computational model is trained to predict a certain unknown output. This training uses a target function represented by a finite set of training examples of inputs and the corresponding desired outputs. At the end of this process, the model should be able to predict correct outputs from their input as well be able to generalize to previously unseen data. Poor generalization can be characterized by over-training. If this happens the model is just memorizing the training examples and unable to give correct outputs from patterns that were not in the training dataset. These two crucial factors (good prediction and good generalization) are conflicting; this problem can also be called the bias and variance dilemma [58].

For this work, the data given to the training step will have initially only information from present time (n=1), and then two steps (n=3) or four steps (n=5) into the past, to test its impact on the model performance.

Another important step is the data splitting for the model to train, validate and test. Using the data in a time sequence, known as systematic sampling, with data for 1 year only the model training

would be limited to only a set of summer (low cloudiness) or winter (high cloudiness). Instead, a convenience sampling algorithm [60] was implemented, making it possible to train, validate and test the model with the information of the full year. Data was divided in 5 days for training (approximately 56% of the data) then 2 days for validation (22%) and 2 days for testing (22%), this process was repeated until the end of the year. The days that could have been left at the end were added to the training set. This process can be seen in Figure 3.2.

Figure 3.2: Process of Selection of days for Training, Validation, and Testing

Each model forecast and clear-sky was restricted to daylight hours since forecasting PV at night is trivial.

Finally, a special case study where all stations selected were summed, creating a case for the regional forecast, was also tested. In this case, all 57 stations plus a new variable that is the sum of all stations in the specific time, will be used as input for the models.

3.2. Clear-sky model

In the context of this work, the Lonij approach was chosen since it is easy to understand and to implement and does not require detailed information from the systems used, only past PV generation records. As mentioned in the previous chapter is a great model to remove the impacts that the position of the system or the season of the year can have in our forecast. Another benefit of using this index is the isolation of the clouds interference since a clear day has values of K=1 and with K near 0, we have very cloudy days.

For this approach we need some information from the past days to create our no "cloud" reading, normally we need between 1 to 2 weeks readings. The reason for this is because the irradiation reading for the same instance each day will be somewhat constant contrary to the presence of the clouds, as such is possible to get days without any clouds in our time chosen, to get our clear sky production. To select this point is chosen a high quantile, near 100% to isolate the clear sky day for the moment needed. The quantile is not 100% since in this manner is possible to eliminate any inflated reading, equation (3.1) [52].

$$Clear\ sky(t) = Perc[\ y_i(t\ in\ nday), 80]\ with\ n \in \{0 \dots past\ days\} \tag{3.1}$$

where $y_i(t)$ is the yield $kW/kWpeak$(PV/rated power) at time t for system i, and *nday* is an integer in the range {0 to past days}.

After forecasting the clear-sky index prediction, it is multiplied by the PV clear-sky expectation to obtain a PV power production forecast, in all forecast.

3.3. Forecasting

In R after every model was created using the *predict ()* function, to get the forecast for K, since this function can take the information from the model created and used it for the forecast, formula (3.2).

$$Forecast = predict(model, variables) \tag{3.2}$$

Each model was trained and tested for forecasts of 1 to 6 hours ahead (horizon =1 to 6) with the information from the present (n=1) or the past (n>1).

The first point of each day will need to use information from the previous day as such we are using a value from several hours apart and possibly values from different times of day (e.g. using night readings to forecast morning values), this could add error to the forecasting model. To avoid this risk, the forecast is done when is possible to use entries of the same day (e.g. for horizon 1 the forecast is done only after the 2º instant, for horizon 2 only after the 3º instance and so on). This limitation could be addressed by adding extra information, e.g. meteorological data, that in this context is not available.

3.3.1. Persistence

Persistence is modeled by equation (3.3), used to predict the m minutes-ahead forecasting (h=30, 60, 120, …min).

$$F(t+m) = F(t) \tag{3.3}$$

where F(t+m) is the forecasted clear sky index at the time (t+m).

Since this model does not need a validation process the data that was divided into two parts, training and testing set, adding the validation set to the training, shown in Figure 3.3.

Figure 3.3: Process rearranged for no Validation Set

3.3.2. Multivariable Regression

The R linear fit function *lm ()* is used as shown in equation (3.4) where y and x are the response and explanatory variables, respectively. This model used the same process, for train and test, of the persistence since it is not required a validation process.

$$ARX\ model = lm(y \sim x, data = index\ K) \tag{3.4}$$

The R function *predict()* can then be used to forecast or other uses such as determining parameters or residuals [59].

3.3.3. Support Vector Regression

SVR was modeled using the R interface libsvm in package e1071, *svm ()*, models are fitted, and new data are predicted as usual, and both the vector/matrix and the formula interface are implemented.

$$SVR\ model = svm(y \sim x\ , data, kernel, type, gamma, cost, epsilon, nu) \tag{3.5}$$

The *kernel* option has 4 choices, linear, polynomial, Gaussian/radial basis (RBF) and sigmoid kernel, described Table 3.1. The linear kernel is a linear model. The polynomial kernel is similar, but the boundary is of somewhat defined but arbitrary order, RBF uses a normal curve around the data points, and sums these so that the decision boundary can be defined by a type of topology condition such as curves where the sum is above a value, the last kernel, sigmoid is like a logistic regression. As stated before choosing the kernel to use is not a trivial manner and still there isn't an

effective way to do, so in this case, for simplicity, was used the most common and most versatile one, RBF.

Table 3.1: Kernels Available

Linear	$\boldsymbol{u'v}$
Polynomial	$(\boldsymbol{\gamma u'v + coef0})^{\boldsymbol{degree}}$
RBF	$\boldsymbol{e^{(-\gamma \lvert u-v \rvert^2)}}$
Sigmoid	$\boldsymbol{tanh^{(\gamma\, u'v+coef0)}}$

The function *svm ()* tries to be smart about the mode to be chosen, using the variable *y*, if *y* is a factor, the engine switches to classification mode, otherwise, it behaves as a regression machine. The *type* option can still be manually chosen for selecting the mode used. In this case, the modes selected were eps-regression and nu-regression.

The *gamma*, *cost*, and *epsilon* are the parameters selected by the user for the eps-regression and RBF kernel, *gamma* being a specific parameter for the kernel function and can be thought of as the "spread" of the kernel, or the decision region.

Table 3.2 presents the values used in the early phase of this work, following [60] that uses a grid search approach to select the parameters to use in their SVR model for solar power forecasting. These values are used as an initial starting point.

Table 3.2: Parameters used for SVR

Cost	0.25, 0.50, 1.0, 2.00, 3.00, 4.00, 5.00, 6.00
Epsilon	0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40
Gamma	7.81×10^{-5}, 1.56×10^{-4}, 3.13×10^{-4}, 6.25×10^{-4}, 1.25×10^{-3}, 2.50×10^{-3}, 5.00×10^{-3}
nu	0.25, 0.50, 0.75

The selection of the parameters (optimization) was done by simplified grid-search, creating a model for each combination of parameters and choosing the one that got the lowest RMSE in the validation set.

In case of the nu-regression, there is a new parameter *nu*, with this variable the value of epsilon is not manually selected since this will be automatically calculated by the model, the other parameters stay the same.

3.3.4. Forecast Accuracy Measures

Forecast accuracy depends on the region, evaluation period, forecast horizon, etc, making comparisons between forecast methods rather challenging. The Agency Solar Heating and Cooling Program Task 36 on "Solar Resource Knowledge Management" and the project "Management and Exploitation of Solar Resource Knowledge" suggest guidelines for benchmarking and conducted comparisons of different solar forecast models against sets of common ground station data [61], [62]. Using the report from Beyer [63], the forecast accuracy for PV production was assessed in terms of RMSE, MAE, BIAS defined in equations (3.6), (3.7), (3.8).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(Prediction, i - Observation, i)^2}{N}} \tag{3.6}$$

$$MAE = \sum_{i=1}^{n} \frac{|Prediction, i - Observation, i|}{N} \tag{3.7}$$

$$BIAS = \sum_{i=1}^{n} \frac{(Prediction, i - Observation, i)}{N} \tag{3.8}$$

The prediction represents the PV power forecast and the observation the real reading in the same instance, where the sums are carried out over all N, that represents the length of the set used.

RMSE gives more weight to large errors, whereas MAE reveals the average magnitude of the error, the BIAS indicates if there is a significant tendency to systematically over-forecast or under-forecast.

A good way to compare models is by selecting one model as a benchmark (normally a simple one). This can be seen done by Beyer [63], who a creates a skill score parameter, in the case of this work will be done the same with the RMSE, this can be defined by equation (3.9).

$$Skill = (1 - \frac{RMSE\ Model\ Tested}{RMSE\ Persistence\ Model}) \times 100 \tag{3.9}$$

This skill equation indicates the percentage improvement in RMSE over the reference model used, in this case, persistence. If a skill score of 100% indicates a perfect forecast improvement, a score of 0% indicates no improvement and a negative skill means that the forecast model shows worse performance over the reference.

Another evaluation procedure consisted of using the normalized RMSE, equation (3.10).

$$nRMSE = \frac{RMSE\ Model}{kWp\ system} \times 100 \tag{3.10}$$

With this method, it is possible to see the weight of the RMSE by the system peak production.

3.4. Computation time management

After the first results of working with SVR, an important issue was noted: the time consumed for the optimization process in the ε-SVR for the 57 stations. This time could go from 26-hours to 156-hours from 1 or 6 horizons, with 30 min step and the parameters in Table 3.2. It is thus unpractical to test the SVR model.

This problem could/should be mitigated by using a faster computer since the computer used for this work was an old Intel® Core™ i3-3217U, a relatively slow processor. Another way to reduce this problem would be to reduce the number of parameters used in ε-SVR and reducing the size of the data given to the creation of the model and stations to predict.

Since the time used for this model was too consuming, trimmed datasets were explored to estimate its impact on computation time using five random cost, epsilon and gamma values. Figure 3.4 shows the results.

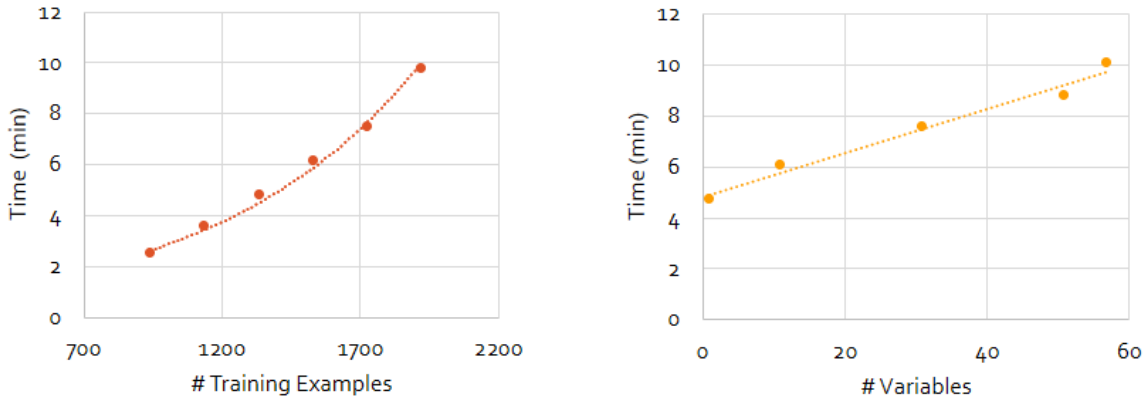

Figure 3.4: ε-SVR time distribution

Increasing training examples is associated with a power trendline tendency while adding variables leads to linear time increase. This means that the numbers of training sets are what will significantly increase or reduce time, not the number of variables used.

The first attempt to reduce the time taken was to transform the step of the data from 30 min to 1-hour. This reduced the training examples in each model by half. Still, the time taken was very demanding and increasing the step more would impose a large limit to the models. So instead of reducing the data more, the forecasting was done to fewer stations.

The stations selected from the 57 were done with a basis in the ARX skill, making rankings of 15, best to worst, over all stations. A representation of this method can be seen Figure 3.5**.**

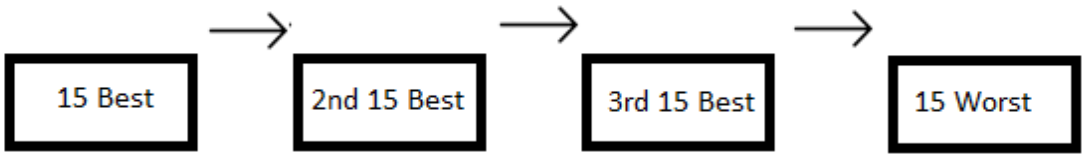

Figure 3.5: Order of Selection for the Systems to be studied

With the ranking done, the top three of the best and bottom three of the worst were selected and in the rankings between, the middle ones were selected, having been chosen 12 of the 57 stations. The stations selected are marked in red in Figure 3.6. Note that all 57 stations are still used as input for the forecast model variables of the 12 stations selected.

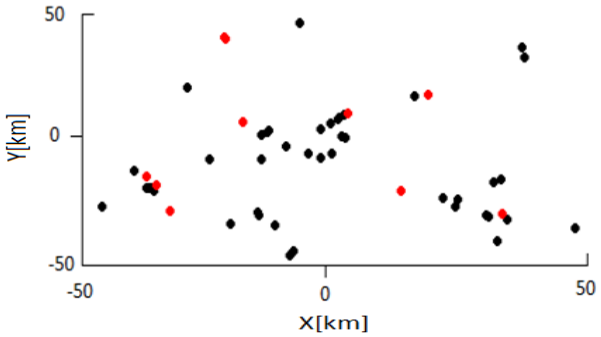

Figure 3.6: Representation of the Systems Chosen

The last approach to decrease computing time was shortening the range of parameters used. For this, a color map for horizon 1 was done, for the best, mid and worst station, Figure 3.7, Figure 3.8 and Figure 3.9.

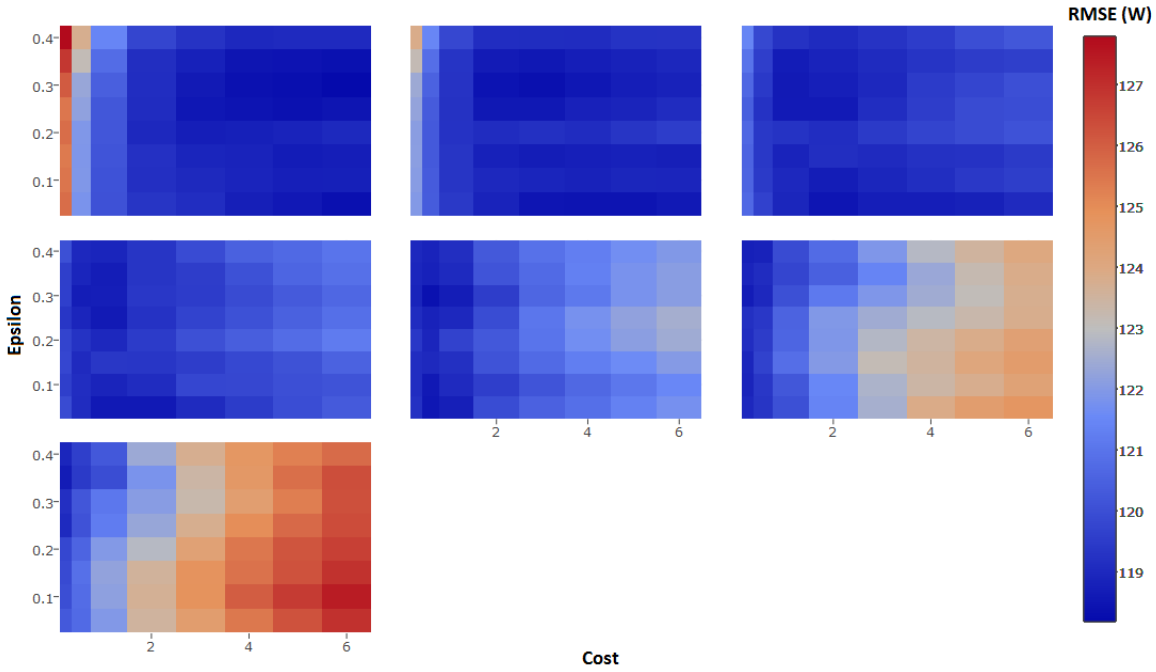

Figure 3.7: Results for a different cost, epsilon and gammas combinations. Each plot is a new gamma. Best RMSE Station

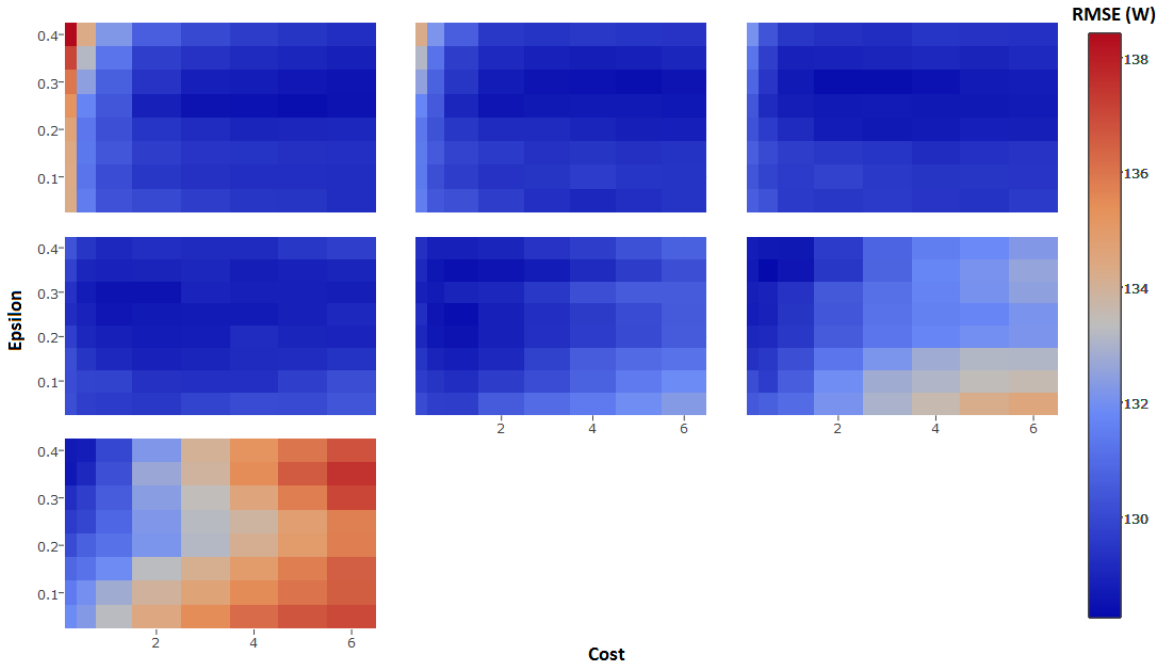

Figure 3.8: Results for a different cost, epsilon and gammas combinations. Each plot is a new gamma. Medium RMSE Station

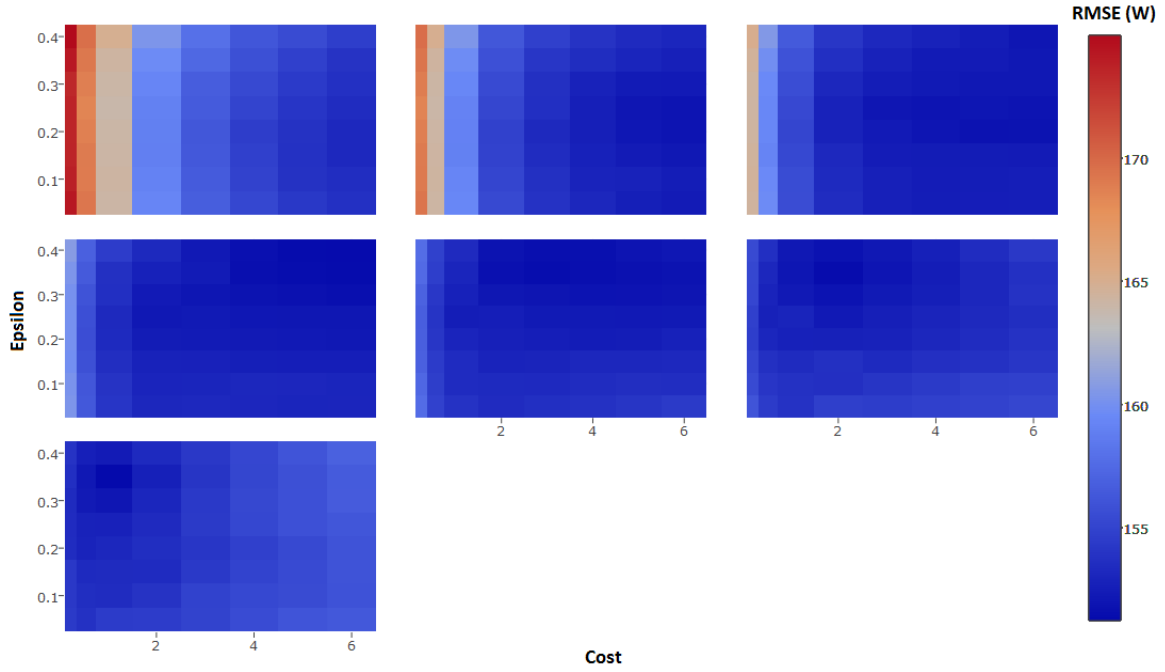

Figure 3.9: Results for a different cost, epsilon and gammas combinations. Each plot is a new gamma. Worst RMSE Station

Each plot, in each figure, represents a gamma used with the order (right to the left) equal to the one in Table 3.2. In this we can see that the RMSE increases with higher gammas, meaning we could remove the biggest gammas, in this case, the gamma 2.50×10^{-3} and 5.00×10^{-3}.

Looking at the cost, we can deduct, like the gamma for higher values, that the RMSE increases with the cost: Hence, the values above 4 were removed. Another remark is for the smaller cost

values, 0.25 and 0.50, where they have a practically equal response in every scenario, being chosen the value 0.50.

The last thing to look is the epsilon value. In the first two gammas, higher values of epsilon have the highest RMSE, being very little difference in the remaining gammas, so the three highest epsilons were removed. After this process, the parameters to work with can be seen in Table 3.3.

Table 3.3: Filter Parameters for SVR

Cost	0.50, 1.0, 2.00, 3.00, 4.00
Epsilon	0.05, 0.10, 0.15, 0.20, 0.25
Gamma	7.81×10^{-5}, 1.56×10^{-4}, 3.13×10^{-4}, 6.25×10^{-4}, 1.25×10^{-3}
nu	0.25, 0.50, 0.75

The data was handled with a step of 1 hour and horizons (forecast) to 6 hours, where only the installed power and the historical generation are known.

These modifications led to a significant reduction of computing time, allowing practical forecast analysis. Results are presented in the next chapter.

4. Results

This chapter presents the results obtained from the different models using the methods described in chapter 3. PV power forecasted by each model are compared with the real data. The starting point is the method using only present information (n=1) followed by the addition farther past information. Is also done a brief comparison with results from other studies, presented in chapter 2.

4.1. Forecasting models results

This section presents the contrast between the forecast and historical readings for some examples (examples were taken from the month of July), a visual representation of the results. Figure 4.1 shows the example forecast for horizon 1 of each model, with the real reading for one individual system with n=1 (this system will be the same in each example).

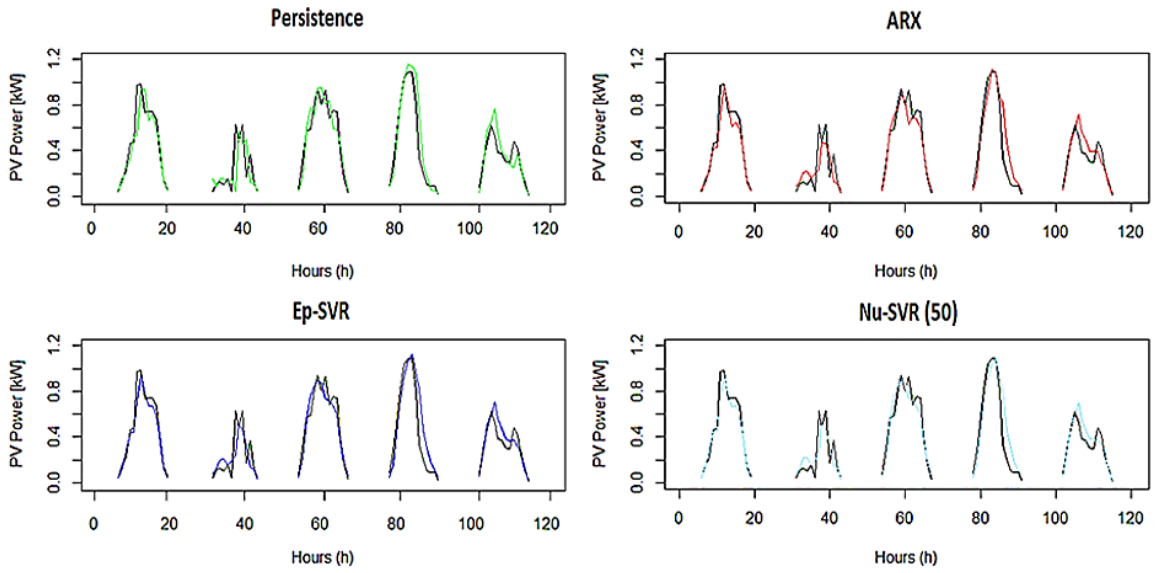

Figure 4.1: PV reading [black line] vs Forecast PV [colour line] for 5 days (Horizon =1, one random individual system)

In this example each model seems to follow the real reading tendencies, not being a “big” difference. Each of them seems to be doing a good work predicting the shadows behavior. In the case of horizon 6, Figure 4.2, some differences are noticeable.

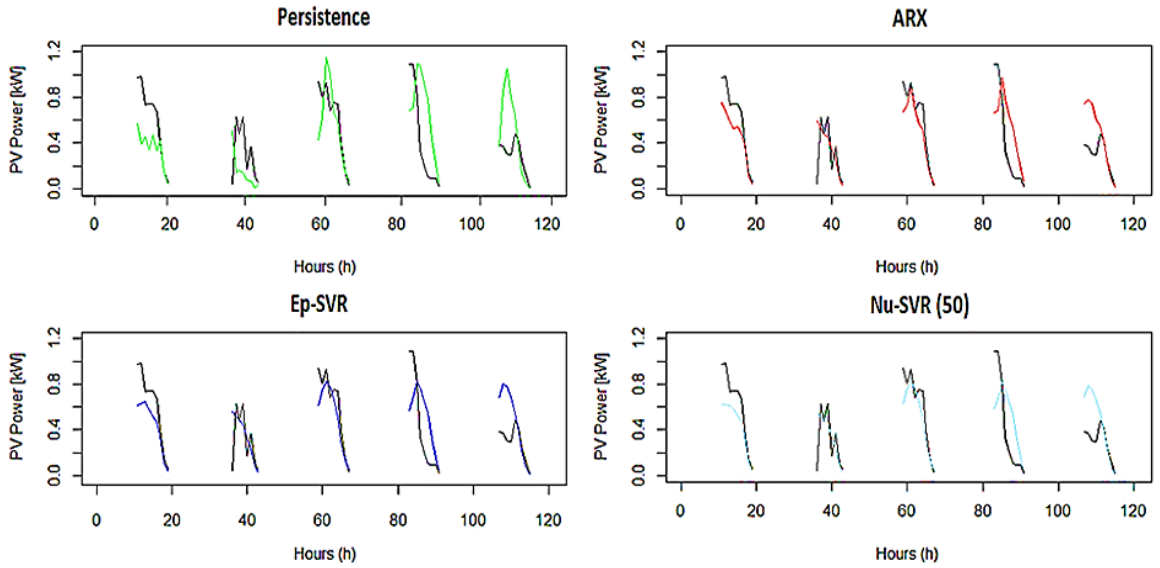

Figure 4.2: PV reading [black line] vs Forecast PV [colour line] for 5 days (Horizon =6, one random individual system)

In this case, we can clearly see that the Persistence does not provide accurate predictions compared to ARX and both SVR's and ARX can be somewhat distinguished, especially in the first and fourth day.

Looking at the regional data prediction for the same days and horizon 1, Figure 4.3.

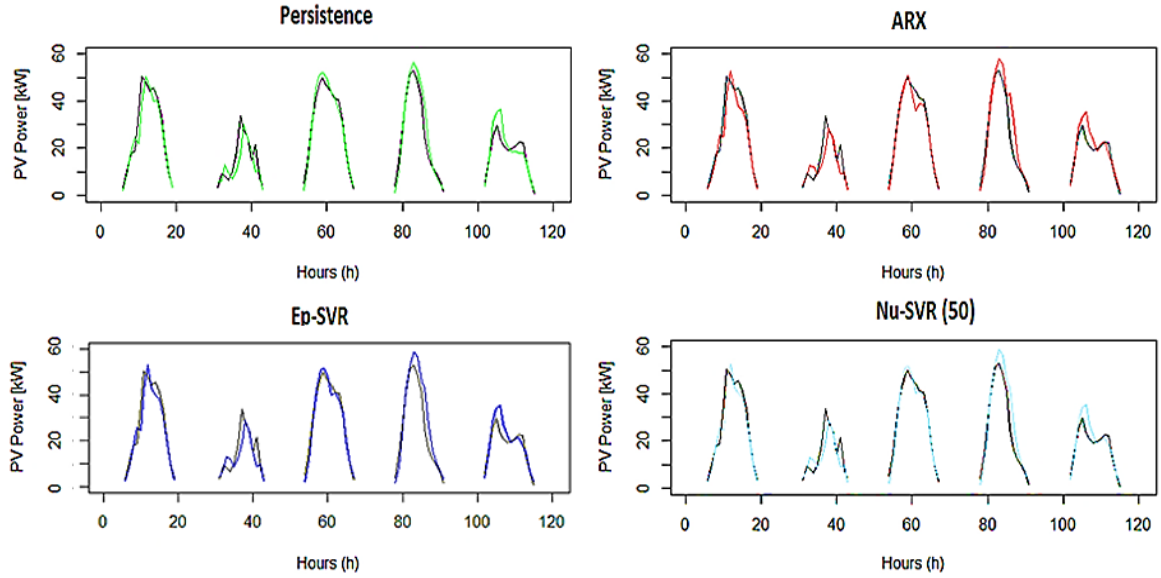

Figure 4.3: PV reading [black line] vs Forecast PV [colour line] for 5 days (Horizon =1, Regional system)

In this example, it is possible to tell that the real readings become smoother and each model seems to have equal performances, as the case with the individuals. Since the real readings are smoother the persistence has less trouble following it as can be seen. The same does not occur for horizon 6, Figure 4.4. Since we are using present information (n=1) to see 6-hours ahead (horizon 6) the persistence shows the vulnerability of having lower performances for higher forecast horizons. The other models show performances like the individual cases.

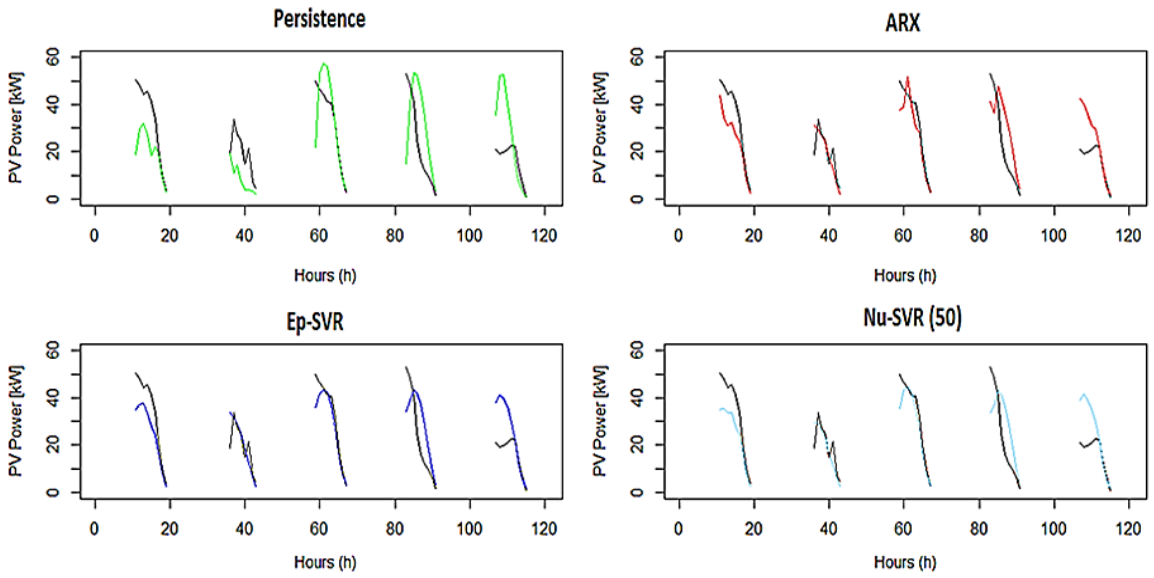

Figure 4.4: PV reading [black line] vs Forecast PV [colour line] for 5 days (Horizon =6, Regional system)

4.2. Comparison of forecasting methods

This section will present the results obtained with the scenario of using only present information (n= 1), the scenarios of adding past information (n=3 and 5) and the results of using ν-SVR.

4.2.1. Present information scenario

Upon the development of the different forecasting models according to the methods described in chapter 3, the models were applied to the selected datasets. In Figure 4.5, shows the BIAS for each model and each station for the 6 horizons and the regional data. At first look at the individual case, is possible to say that the ARX and SVR models tend to under-predict as the horizon grows, having a BIAS much less variable than the Persistence. The Persistence case, presents a very variable BIAS easily distinguished beyond horizon 2. This means that the model has a low accuracy, contrary to the ARX and SVR case.

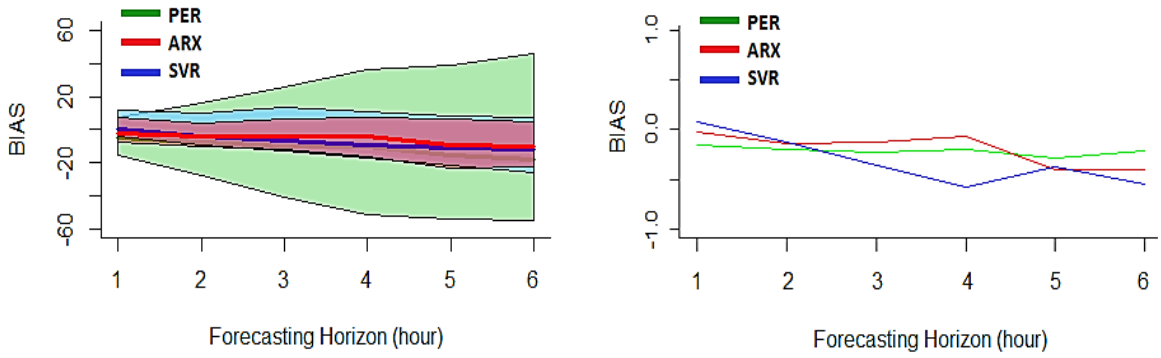

Figure 4.5: BIAS Results (Left: Individual System, Right: Regional System) Note: The areas shown in the left plot correspond to the ranges for each model considering the highest and lowest system's BIAS value

Regarding the regional data, the Persistence model features a constant BIAS near zero regardless of the horizon. This is expected, since aggregating the data will smooth out the production curve

hence increasing the accuracy of the Persistence. The other models still have the tendency to under-predict like the individual case.

Looking at the MAE, Figure 4.6, the first insight is the Persistence model, where it features a higher MAE overall than ARX and SVR; this effect is accentuated for longer horizons. The same can be said for the regional data.

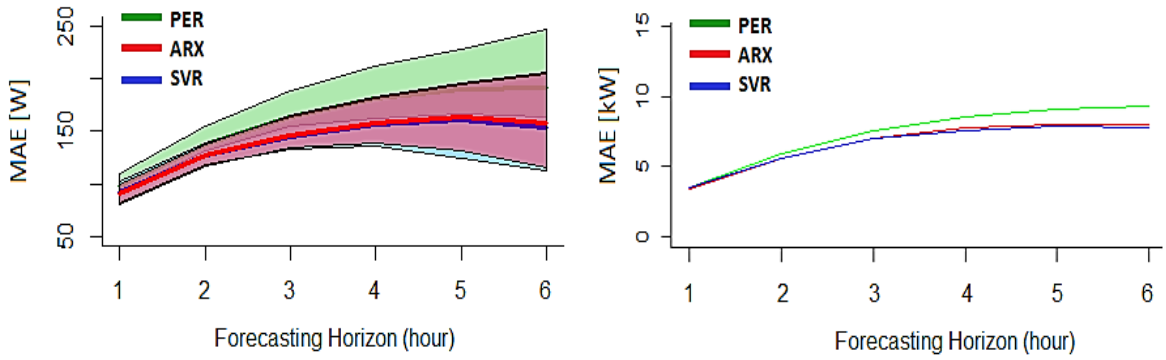

Figure 4.6: MAE Results (Left: Individual System, Right: Regional System) Note: The areas shown in the left plot correspond to the ranges for each model considering the highest and lowest system's MAE value

The improvements of the ARX and SVR methods over Persistence are reduced which suggest that errors are due to miss-timing passing clouds. Indeed, a passing cloud leads to a strong variation in the PV output and therefore miss-timing its arrival at the target location leads to a higher MAE error.

Regarding the RMSE, Figure 4.7, where it gives more weight too larger errors, is more evident the behavior reported before, in Figure 4.6. The Persistence now is clearly with the worst performance with ARX and SVR presenting values nearly equal, with advantage for SVR.

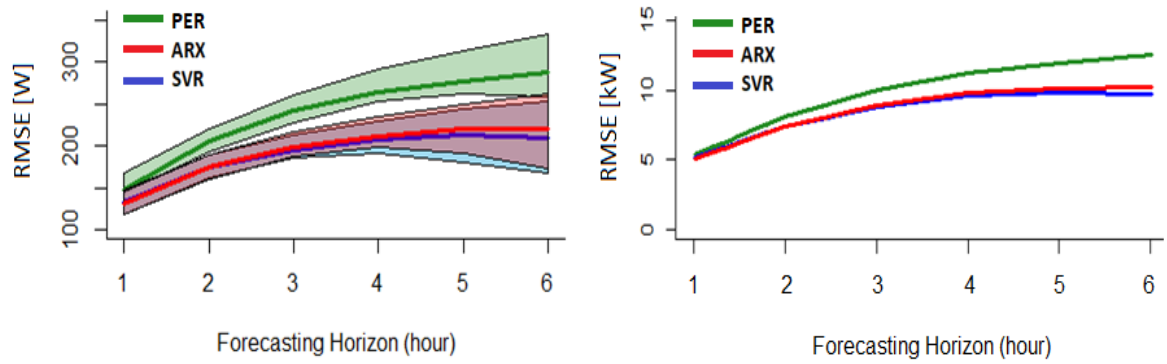

Figure 4.7: RMSE Results (Left: Individual System, Right: Regional System) Note: The areas shown in the left plot correspond to the ranges for each model considering the highest and lowest system's RMSE value

Using the normalized RMSE, we can also see the impact it has on the peak power. Like Figure 4.7, the Persistence as the worst performance, being ARX and SVR very close to each other. The impact we can have in the peak power can vary for near 6% to 10%, Persistence case, or 4% to 7%, ARX and SVR, from horizon 1 to 6., with best performances for the SVR case.

The RMSE can also be presented as forecasting skill, Figure 4.8, using Persistence as a reference for a better and easier comparison. In both cases, it shows improvements that can go up to 25% for ARX or 30% for SVR in some stations (h=6). For the individual case, ARX and SVR have a

practically equal behavior, the main difference is that SVR demonstrates better forecasting skill in general than the ARX, even if the median is practically equal.

In the regional case can be seen that skill is much lower than the individual case (nearly half). As discussed above, this effect can be mainly attributed to the improvement of the aggregated regional Persistence. Even so, for the farthest horizons, the improvement is practically equal to the median of the individual cases.

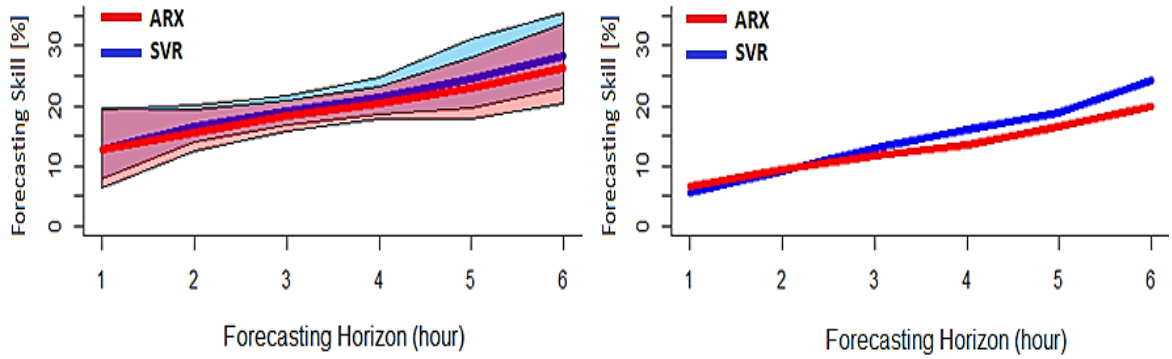

Figure 4.8: Forecasting Skill [%] (Left: Individual System, Right: Regional System) Note: The areas shown in the left plot correspond to the ranges for each model considering the highest and lowest system's Skill value

In Figure 4.9, can be seen the effect of adding the azimuth and the solar angle (Astro variables) to the model. It shows that adding these new variables had nearly zero impact for the models, for both individual and accumulated data.

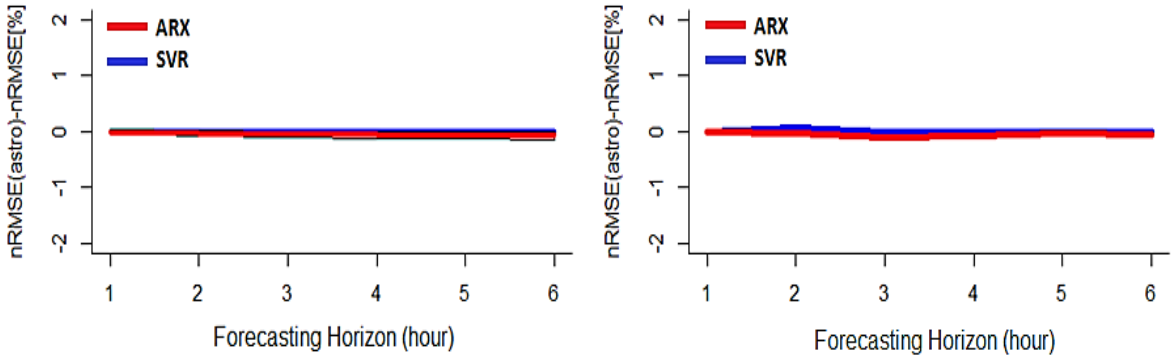

Figure 4.9: Adding Azimuth and Solar Angle (Left: Individual System, Right: Regional System)

4.2.2. Adding Past Information

As discussed in chapter 3, data section, with basis the article from Kariniotakis and co-authors [57], the forecasts may be improved by using past information and not only present generation data. This is relevant when the forecast horizon is of the order of magnitude of the time step and the time clouds take to travel from shadowing a neighboring station to shadowing the target station.

Figure 4.10, using n=3, SVR and ARX start to show noticeable differences, especially beyond horizon 3. This figure ought to be compared with Figure 4.8 above, both still demonstrate improvements over Persistence but the ARX has lower skills than the SVR. Both models improve the skill as the horizon increases. As for the regional data, the difference between ARX and SVR

only starts to be noticeable after horizon 4, being SVR nearly 8% better in horizon 6. On the horizon of 1-hour ARX demonstrates slightly better results (1%).

When adding more 4 points of past information (n=5) we can clearly see that the ARX model, after horizon 4 performs much worse than SVR and even well below persistence, achieved considerably negative skills, for both individual or regional case. The SVR still presents skills always positive, with a tendency for improvement.

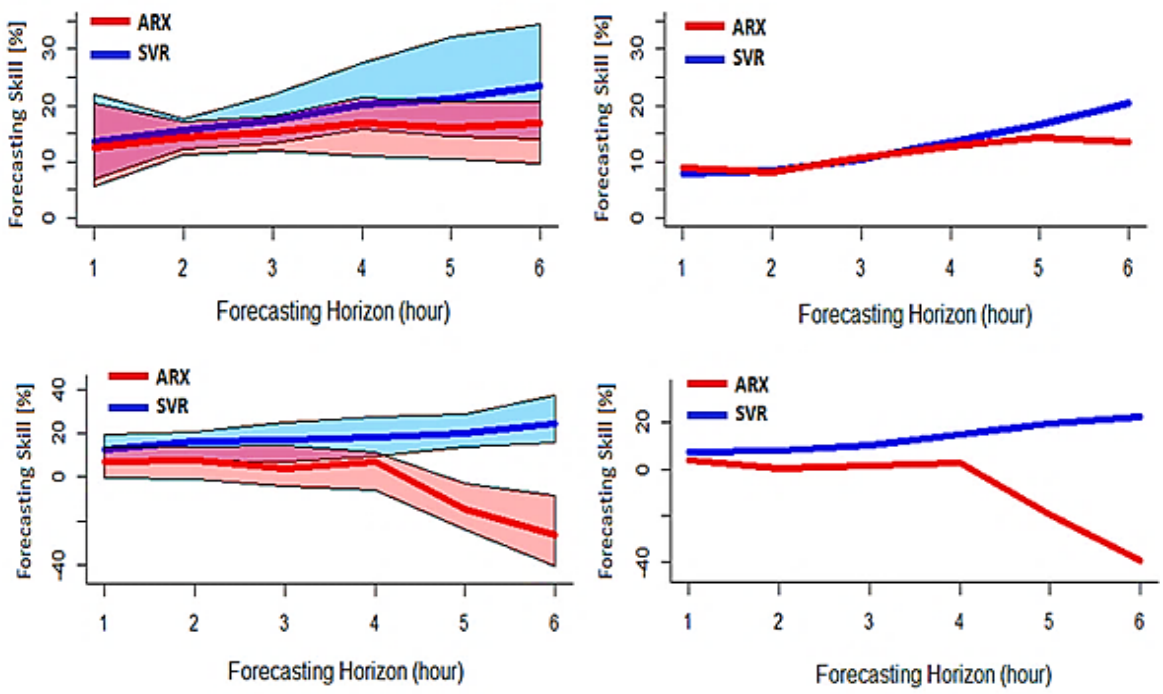

Figure 4.10: Results for Skill [%] with n=3 (top) and n=5 (below) (Left: Individual System, Right: Regional System)
Note: The areas shown in the left plot correspond to the ranges for each model considering the highest and lowest system's Skill value

The impacts are not that noticeable for the nRMSE for n=3, in Figure 4.11, all models still have the same tendency as the one for n=1. The biggest difference is in the intervals from each model, becoming larger, meaning, a higher variability of RMSE for each station. For the regional case, it shows that adding this extra past information, had good results, decreasing the nRMSE for each model, after horizon 3. On the n=5 scenario after horizon 4, ARX has a downgrade in performance, easily noticeable from the other models.

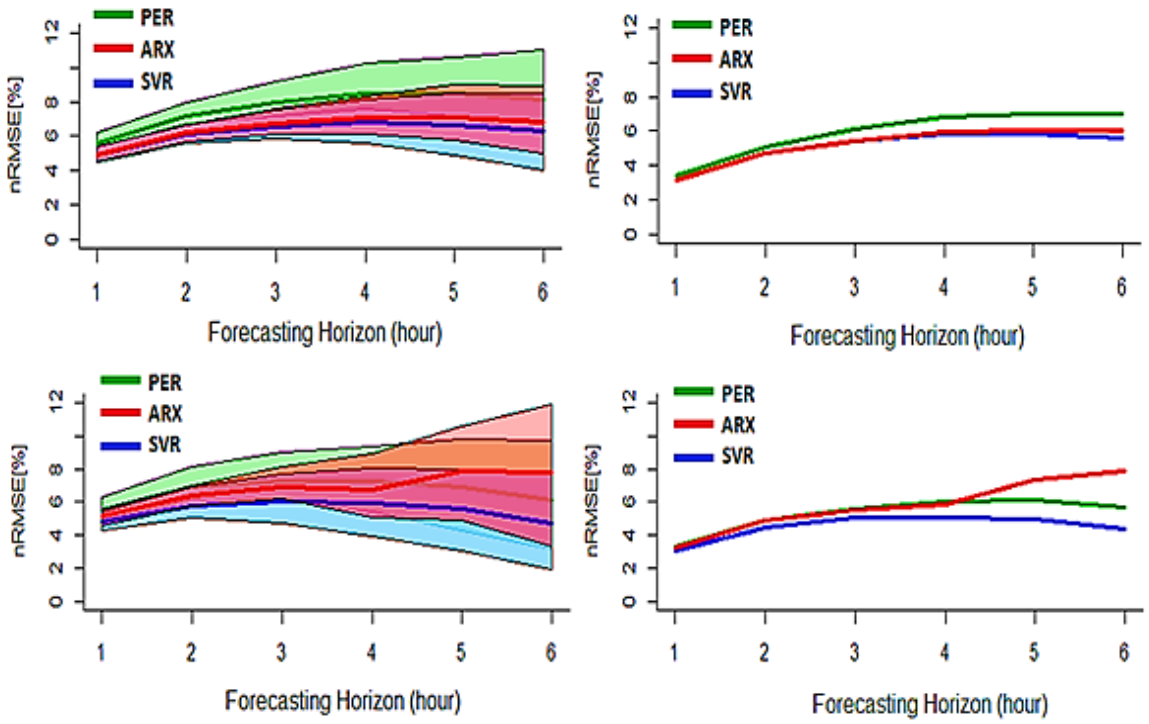

Figure 4.11: Results for nRMSE [%] for n=3 (top) n=5 (below) ((Left: Individual System, Right: Regional System) Note: The areas shown in the left plot correspond to the ranges for each model considering the highest and lowest system's nRMSE value

This worsening of the ARX can be mainly caused by the reduced number of targets for the model to train. Since adding more past information with the methodology used for removing the night will automatically remove the forecasting for the first n and plus horizon points of each day (only forecasting afterward). This is probably making the ARX overfit over the training set and losing his capacity to generalize well to the new data. Figure 4.12 shows the effects of this reduction on the test set, seeing that increasing n =1 to 5 will reduce the data to forecast in approximately 33%. This reduction is more impactful for higher horizons and for lower resolution data (for example is worse using 1-hour data than using 30 min data). On the other hand, it seems the SVR model can handle very well this reduction and still presents reasonable performance.

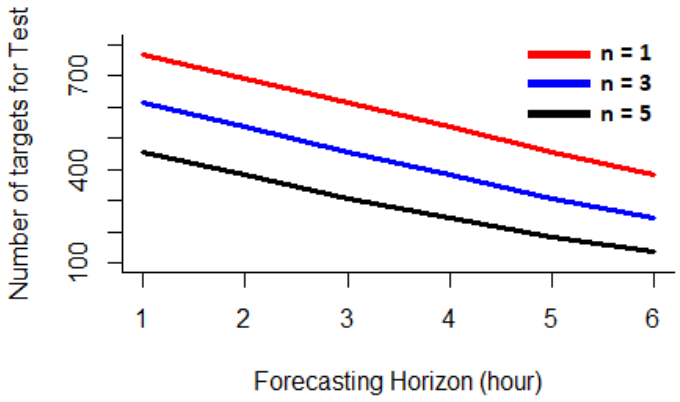

Figure 4.12: Number of points forecasted in the test set for every model

4.2.3. ν-SVR

An alternative method to reduce the computation time consumption is ν-SVR since we do not need to optimize the epsilon. The counterpart is optimizing the variable nu, which is somewhat easier. Using the three-selected nu (0.25, 0.50, 0.75) the time consumed was reduced from 82% (nu=0.75) to 91% (nu=0.50). Figure 4.13 demonstrates the impact of skill each nu had.

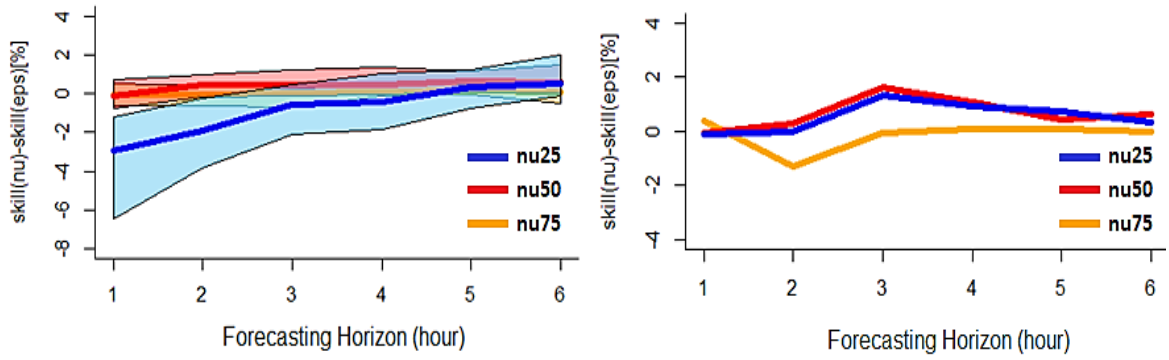

Figure 4.13: Results of Skill [%] using different nu to the ϵ-SVR counterpart (Left: Individual System, Right: A cumulative System) Note: The individual case stain range of each model is the interval between the highest and lowest system's Skill value

First thing noticed is about ν 0.25, in horizon 1 can have a reduction in the skill of 6% to 2% (individual case), not being very viable, even if it would reduce the time in 91%, from ε-SVR. This is compensated for larger horizons (5 to 6) that the lost skill is near 0% (median). The other nu (0.50 and 0,75) have a similar behavior (almost undistinguished from each other) and nearly 0% losses in skill. The point that highlights them is the time consumed, a reduction of 85% for ν 0.50 and 82% for 0.75. The same idea can be taken for the regional data, having fewer losses in skill.

The good particularity of this test with ν-SVR is, contrary to expectations, in some cases, it could enhance the skill forecast. This can particularly be easily seen in the regional case for horizon 3. This could be explained by the automatic optimization process for epsilon in ν-SVR was better than the epsilon range (chosen manually) used, for the optimization in ε-SVR.

Comparing with nRMSE, Figure 4.14 shows that the losses for each nu is nearly 0% and ν 0.25 can be a little better. But this difference is insignificant and can be said that they had zero impact on the nRMSE.

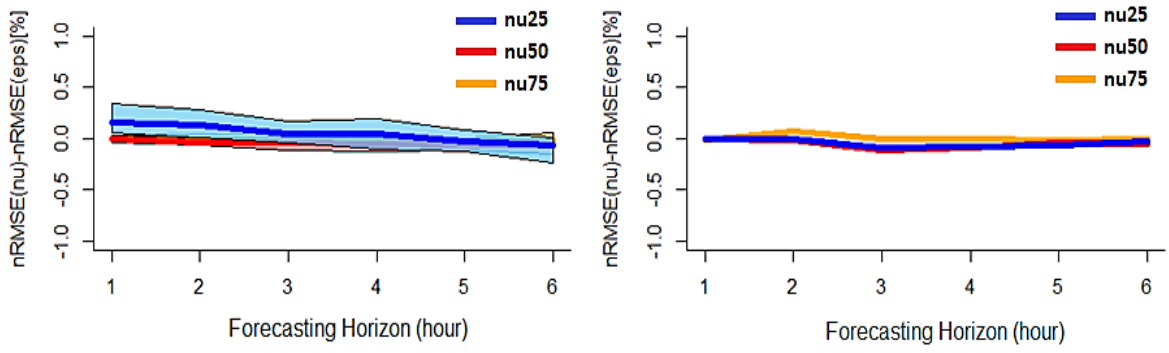

Figure 4.14: Results of nRMSE [%] using different nu to the ϵ-SVR counterpart (Left: Individual System, Right: A cumulative System) Note: The individual case stain range of each model is the interval between the highest and lowest system's nRMSE value

4.3. Benchmark analysis

Is hard to compare results in numbers of RMSE, nRMSE or MAE with those in the studies presented in the studies overview section, in chapter 2, since everyone used different data from the different location. But the overall results main ideas can be compared. The study from Sharma, [54], that used SVM-RBF with mainly focused in the correlation between the variables (like humidity and precipitation) had the result that could outperform more simple models in 27%, the results of this work reinforced this idea with the constant SVR outperform over the Persistence model.

In Xu and co-workers, [56], work, using WSVM, showed that adding more information to the model (in Xu work the information added were more similar days history) had a negative impact in performance. This work showed that adding more information to the training set of the model dint had a positive impact also, but this scenario needed more study and a good step to take would be using the data in 30 min resolution and not in 1-hour and see if this would help the model. On the other hand, this model can work well with the lack of information.

In overall, this work reinforces the potential that SVR has for forecasting PV power, like what Pelland and co-authors, [4], said.

In all the cases studies, both ARX and SVR outperform the Persistence method, with the exception in the case using 4 past point where ARX for larger horizons had worst results. The results demonstrated that using information from nearby stations can help the previsions accuracy. Specifying the case using only present data for the forecast, both more complex models (ARX and SVR) had a very similar performance, with advantage for SVR in higher horizons, this tendency is equal for both individual and aggregated data. A feature in the accumulated data prevision is to have lower skill than the individual case for lower horizons but being nearly equal for larger horizons. This can be explained by the better performance than the Persistence model had for the accumulated reading.

5. Conclusions

The work done in this study deals with solar power forecasting based on multivariable linear regressions (ARX) and support vector regressions (SVR) for a set spatially distributed photovoltaic (PV) system's, and their aggregate, using a data-driven clear sky index for the multiple neighbor PV systems. The Persistence model was also used as the base of comparison between ARX and SVR model.

The data used for testing and construction of the R script was originated from a region in the United Kingdom, that contained information from one year of PV production (01/07/2015 to 30/06/2016), with a 30 min step. A preliminary analysis identified the PV systems with suitable data, eliminating those locations with too many missing values.

The Persistence model was the first forecasting model to be tested, as it is a good base for comparison to the other models. The first models only consider present information (n=1). Further models include farther past information, up to 4 hours.

The first runs of the SVR model showed unpractical time taken for the optimization process in the ε-SVR for the 57 stations, from 26-hours to 156-hours depending on the number of horizons considered. This problem was mitigated by reducing the number of parameters used in ε-SVR and reducing the size of the data set by decreasing the number of target locations. An alternative option, ν-SVR, was also tested, further reducing the time consumption problem.

ARX and SVR outperformed the Persistence model in all tested conditions apart from the case were 4 past points were used (n=5), In this case, ARX seems to fail for larger horizons. The overall results demonstrated that using information from nearby stations can help the forecast accuracy.

When past information is added (n=3 and n=5) SVR outperforms ARX. In general, both models performed worse than only using present variables (n=1). This means the past information not only does not help the forecasting but makes it worst. This might be caused by the increase of inputs and the reduction of the targets for the model, making the ARX overfit the training set and losing his capacity to generalize well to the new data. This effect might be limited if the 30 min step data was used, as it would keep enough targets for generalization. One important conclusion is that the SVR model is a more robust model than the ARX and can better handle more variables and fewer targets, not suffering from over-fitting the predictions, that seems to be the main issue with the ARX model.

The SVR model is very computing demanding, a lot more than the ARX. In ε-SVR there is no standard procedure for parameters selection, requiring a trial and error approach. The alternative use of ν-SVR considerably streamlines this process without significative loss of performance.

For short horizons (between 1 and 3 hours) all models tested produce reliable results. For larger horizons (from 4 hours onwards) the SVR model is clearly the best option.

A combination of these models could be a good alternative to the trade of time taken and performance wanted. For example, we could implement the ARX model for the morning and late afternoon, since error in this situation does not have a big impact in the PV production and implement the SVR for the time between (mid-day), were the solar production is a lot bigger and the error here have a lot more meaning. These models should still be tested with better machines (for example an Intel core i7) to see if the computation time decreases and using better resolution

data should improve both models, especially the SVR one since it appears can overcome the overfitting problem.

Overall, the usage of neighbor's system PV production coupled with SVR and the clear sky index proves to be a serious alternative for providing accurate PV power forecasts, mainly in situations where there is not much information about the PV system.

6. References

[1] T. M. Razykov, C. S. Ferekides, D. Morel, E. Stefanakos, H. S. Ullal, and H. M. Upadhyaya, "Solar photovoltaic electricity: Current status and future prospects," *Sol. Energy*, vol. 85, no. 8, pp. 1580–1608, 2011.

[2] G. K. Singh, "Solar power generation by PV (photovoltaic) technology: A review," *Energy*, vol. 53, pp. 1–13, 2013.

[3] BP, "BP Statistical Review of World Energy 2017," *Br. Pet.*, no. 66, pp. 1–52, 2017.

[4] S. Pelland, J. Remund, J. Kleissl, T. Oozeki, and K. De Brabandere, "Photovoltaic and Solar Forecasting: State of the Art," *Int. Energy Agency Photovolt. Power Syst. Program. Rep. IEA PVPS T14*, pp. 1–40, 2013.

[5] "Solar Server Global Solar Industry Website." [Online]. Available: https://www.solarserver.com/. [Accessed: 01-Jan-2017].

[6] International Energy Agency Photovoltaic Power Systems Programme, "Snapshot of global photovoltaic markets 2016," pp. 1–16, 2017.

[7] M. Lave *et al.*, "Ota City : Characterizing Output Variability from 553 Homes with Residential PV Systems on a Distribution Feeder," *2012 Twenty-Seventh Annu. IEEE Appl. Power Electron. Conf. Expo.*, vol. 29, no. February, pp. 1–3, 2011.

[8] E. Cuce, P. M. Cuce, and T. Bali, "An experimental analysis of illumination intensity and temperature dependency of photovoltaic cell parameters," *Appl. Energy*, vol. 111, pp. 374–382, 2013.

[9] D. M. Tobnaghi, R. Madatov, and D. Naderi, "The Effect of Temperature on Electrical Parameters of Solar Cells," *ISSN Int. J. Adv. Res. Electr. Electron. Instrum. Eng. ISO Certif. Organ.*, vol. 3297, no. 12, pp. 2320–3765, 2007.

[10] M. S. ElNozahy and M. M. A. Salama, "Technical impacts of grid-connected photovoltaic systems on electrical networks—A review," *J. Renew. Sustain. Energy*, vol. 5, no. 3, p. 32702, 2013.

[11] M. T. and D. G. Infield, "Impact of widespread photovoltaics generation on distribution systems," *Renew. Power Gener. IET*, vol. 1, no. 1, pp. 10–16, 2007.

[12] K. Hao, S. Achanta, B. Rowland, and A. Kivi, "Mitigating the impacts of photovoltaics on the power system," *2016 Saudi Arab. Smart Grid Conf. SASG 2016*, no. August 2016, 2017.

[13] N. Miller and Z. Ye, "Report on Distributed Generation Penetration Study," *Contract*, no. August, pp. 1–108, 2003.

[14] E. Liu and J. Bebic, "Distribution System Voltage Performance Analysis for High-Penetration Photovoltaics Distribution System Voltage Performance Analysis for High-Penetration Photovoltaics," *Natl. Renew. Energy Lab.*, no. February 2008.

[15] M. Makhlouf, F. Messai, K. Nabti, and H. Benalla, "Modeling and simulation of grid-connected photovoltaic distributed generation system," in *2012 First International Conference on Renewable Energies and Vehicular Technology*, 2012, pp. 187–193.

[16] S. Cobben, B. Gaiddon, and H. Laukamp, “Impact of Photovoltaic Generation on Power Quality in Urban Areas with High PV Population - Results from Monitoring Campaigns,” *PV Upscale (Intelligent Energy, Eur.*, p. 53, 2008.

[17] G. P. V Systems *et al.*, “Wavelet-Based Islanding Detection in Grid-Connected PV Systems,” *IEEE Trans. Ind. Electron.*, vol. 56, no. 11, pp. 4445–4455, 2009.

[18] European Photovoltaic Industry Association, “Connecting the Sun: Solar photovoltaics on the road to large-scale grid integration.”

[19] B. Espinar, J. L. Aznarte, R. Girard, a. M. Moussa, and G. Kariniotakis, “Photovoltaic Forecasting: A state of the art,” *5th Eur. PV-Hybrid Mini-Gird Conf.*, vol. 33, pp. 250–255, 2010.

[20] Direcção Geral de Energia e Geologia, “Renováveis - Estatísticas Rápidas - Setembro 2017,” 2017.

[21] IHS Markit, “90 GW residential solar by 2021,” 2017. [Online]. Available: http://www.pveurope.eu/News/Markets-Money/90-GW-residential-solar-by-2021. [Accessed: 20-Nov-2017].

[22] M. Schmela, “Global Market Outlook For Solar Power/2017-2021,” *SolarPower Eur.*, p. 58, 2017.

[23] J. Weniger, J. Bergner, and V. Quaschning, “Integration of PV power and load forecasts into the operation of residential PV battery systems,” *4th Sol. Integr. Work.*, pp. 383–390, 2014.

[24] V. Pratap, K. Vaibhav, and D. K. Chaturvedi, “Solar Power Forecasting Modeling Using Soft Computing Approach.”

[25] D. Mayer, L. Wald, Y. Poissant, and S. Pelland, “Performance prediction of grid-connected photovoltaic systems using remote sensing,” *Rep. IEA-PVPS T2-072008*, 2008.

[26] G. R. T. Esteves, B. Q. Bastos, F. L. Cyrino, R. F. Calili, and R. C. Souza, “Long-term electricity forecast: A systematic review,” *Procedia Comput. Sci.*, vol. 55, no. Itqm, pp. 549–558, 2015.

[27] S. Akhwanzada and R. bin M. Tahar, “Long-term Electricity Forecasting: A System Dynamics Approach,” *Iceeb*, vol. 33, no. Iceeb, pp. 5–6, 2012.

[28] B. Dergisi and E. Sciences, “Long-Term Electricity Demand Forecasting : an Alternative Approach With Support Vector,” pp. 45–53, 2010.

[29] V. Kostylev and A. Pavlovski, “Solar Power Forecasting Performance - Towards Industry Standards,” *1st Int. Work. Integr. Sol. Power into Power Syst. Aarhus, Denmark*, 2011.

[30] L. Elke, S. Thomas, H. Johannes, H. Detlev, and K. Christian, “Regional PV power prediction for improved grid integration,” *EU PVSEC WCPEC-5*, no. 0 September 2010, 2010.

[31] P. Mathiesen and J. Kleissl, “Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States,” *Sol. Energy*, vol. 85, no. 5, pp. 967–977, 2011.

[32] P. Sophie, G. George, and K. George, “Solar and photovoltaic forecasting through post-

processing of the Global Environmental Multiscale numerical weather prediction model," *Prog. Photovoltaics Res. Appl.*, p. 13, 2007.

[33] E. Ogliari, F. Grimaccia, S. Leva, and M. Mussetta, "Hybrid predictive models for accurate forecasting in PV systems," *Energies*, vol. 6, no. 4, pp. 1918–1929, 2013.

[34] M. Abuella and B. Chowdhury, "Solar power probabilistic forecasting by using multiple linear regression analysis," *SoutheastCon 2015*, no. October, pp. 1–5, 2015.

[35] D. C. Pattie and J. Snyder, "Using a neural network to forecast visitor behavior," *Ann. Tour. Res.*, vol. 23, no. 1, pp. 151–164, 1996.

[36] C.~Cortes, C.~Cortes, V.~Vapnik, and V.~Vapnik, "Support Vector Networks," *Mach. Learn.*, vol. 20, no. 3, p. 273~-~297, 1995.

[37] R. Huang, T. Huang, R. Gadh, and N. Li, "Solar generation prediction using the ARMA model in a laboratory-level micro-grid," *2012 IEEE 3rd Int. Conf. Smart Grid Commun. SmartGridComm 2012*, pp. 528–533, 2012.

[38] J. Marquez, "Time series analysis: James D. Hamilton, 1994, (Princeton University Press, Princeton, NJ), 799 pp., US $55.00, ISBN 0-691-04289-6," *Int. J. Forecast.*, vol. 11, no. 3, pp. 494–495, 1995.

[39] D. P. Larson, L. Nonnenmacher, and C. F. M. Coimbra, "Day-ahead forecasting of solar power output from photovoltaic plants in the American Southwest," *Renew. Energy*, vol. 91, pp. 11–20, 2016.

[40] R. Huang, T. Huang, R. Gadh, and N. Li, "Solar generation prediction using the ARMA model in a laboratory-level micro-grid," *2012 IEEE 3rd Int. Conf. Smart Grid Commun. SmartGridComm 2012*, no. November, pp. 528–533, 2012.

[41] G. James, DanielaWitten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 64, no. 9–12. 2007.

[42] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de-Pison, and F. Antonanzas-Torres, "Review of photovoltaic power forecasting," *Sol. Energy*, vol. 136, pp. 78–111, 2016.

[43] P. J. Brockwell and R. A. Davis, *Springer Series in Statistics Springer Series in Statistics*. 1997.

[44] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Stat. Comput.*, vol. 14, no. 3, pp. 199–222, 2004.

[45] A. Ihler, "Support Vector Machines (3): Kernels." [Online]. Available: https://www.youtube.com/watch?v=OmTu0fqUsQk. [Accessed: 29-Dec-2017].

[46] C. Chang, C. Lin, and I. Engineering, "Training ν-Support Vector Regression: Theory and Algorithms," vol. 1.

[47] "Lab 1: Solar Radiation & Seasons." [Online]. Available: http://sites.gsu.edu/geog1112/solar-radiation-seasons/.

[48] J. N. Black, C. W. Bonython, and J. A. Prescott, "Solar radiation and the duration of sunshine," *Q. J. R. Meteorol. Soc.*, vol. 80, no. 344, pp. 231–235, 1954.

[49] B. Y. H. Liu and R. C. Jordan, "A Rational Procedure for Predicting The Long-Term Average Performance of Flat-Plate Solar-Energy Collectors," *Sol. Energy*, vol. 7, no. 2, pp. 53–74, 1963.

[50] A. Golnas, J. Bryan, R. Wimbrow, C. Hansen, and S. Voss, "Performance assessment without pyranometers: Predicting energy output based on historical correlation," in *2011 37th IEEE Photovoltaic Specialists Conference*, 2011, pp. 2006–2010.

[51] N. . Engerer and F.P.Mills, "A Clear-Sky Index for Photovoltaics N.," pp. 1–8, 2015.

[52] V. P. Lonij, A. E. Brooks, K. Koch, and A. D. Cronin, "Analysis of 80 rooftop PV systems in the Tucson, AZ area," *Conf. Rec. IEEE Photovolt. Spec. Conf.*, pp. 549–553, 2012.

[53] P. Bacher, H. Madsen, and H. A. Nielsen, "Online short-term solar power forecasting," *Sol. Energy*, vol. 83, no. 10, pp. 1772–1783, 2009.

[54] N. Sharma, P. Sharma, D. Irwin, and P. Shenoy, "Predicting Solar Generation from Weather Forecasts Using Machine Learning," pp. 528–533, 2011.

[55] S. H. Oudjana, A. Hellal, and I. H. Mahammed, "Power Forecasting of Photovoltaic Generation," *Int. J. Electr. Comput. Electron. Commun. Eng.*, vol. 7, no. 6, pp. 334–338, 2013.

[56] R. Xu, H. Chen, and X. Sun, "Short-term photovoltaic power forecasting with weighted support vector machine," *Autom. Logist. (ICAL), 2012 IEEE Int. Conf.*, no. August, pp. 248–253, 2012.

[57] G. Kariniotakis, A. Michiorri, R. Girard, and A. Bossavy, "The impact of available data history on the performance of photovoltaic generation forecasting models," *22nd Int. Conf. Exhib. Electr. Distrib. (CIRED 2013)*, pp. 0856–0856, 2013.

[58] I. Kononenko and M. Kukar, *Machine Learning and Data Mining: Introduction to Principles and Algorithms*. Horwood Publishing Limited, 2007.

[59] "Using R for Linear Regression," p. 9, 2013.

[60] M. Abuella and B. Chowdhury, "Solar Power Forecasting Using Support," no. October 2016, 2017.

[61] E. Lorenz *et al.*, "Benchmarking of different approaches to forecast solar irradiance," *24th Eur. Photovolt. Sol. energy Conf.*, pp. 25–34, 2009.

[62] J. Mailhot *et al.*, "The 15-km version of the Canadian regional forecast system," *Atmosphere-Ocean*, vol. 44, no. 2, pp. 133–149, 2006.

[63] H. Beyer *et al.*, "Report on Benchmarking of Radiation Products," *Report*, no. January, pp. 70–74, 2009.