FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

WindDesign

João da Silva Carvalho

U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Carlos Soares

Second Supervisor: Dalila Fontes

August 1, 2014

WindDesign

João da Silva Carvalho

Mestrado Integrado em Engenharia Informática e Computação

August 1, 2014

Abstract

Wind farms are composed by a group of turbines connected to a network of electric energy. In the last five years the global capacity of wind energy has tripled. In the last year alone the wind energy industry has increased 12,4 per cent in capacity, globally.

Energy generated by wind is quite consistent in long-term, but has significant variations in short-term. Meteorologic studies assists the adjustments of wind networks accordingly to predicted variations. One of the factors that affect the production of energy is the Wake Effect, that consists of the alterations that the passage through the wind turbines provokes on the air flux. Those alterations have impact in the turbine production where that air flows next. Thus, besides the geographic location of the park, the location of the turbines within the park affects its production of energy.

The optimization of wind farms may allow its increase of energy production. To do so, techniques that allows to discover the optimal location of the turbines to maximize the produced energy are used. The goal of this project is to develop a model to optimize the layout of a wind farm based on the expected impact of the turbines locations in the production. To do this data mining and machine learning techniques will be used in order to acquire a model that would predict the production of a turbine based on its location as well as other external variables.

Resumo

Os parques éolicos são constituídos por um conjuntos de aerogeradores ligados a uma rede de transmissão de energia eléctrica. Nos últimos cinco anos a capacidade instalada global da energia eólica triplicou. Só no último ano a indústria da energia eólica adicionou 12,4 por cento de capacidade, globalmente.

A energia gerada do vento é bastante consistente em períodos longos (por ex., anuais) mas tem variações significativas em períodos curtos. O estudo da meteorologia auxilia o ajustamento de redes eólicas de acordo com as variações previstas. Um dos fatores que afeta a produção de energia é o wake effect, que consiste nas alterações que a passagem pelos aerogeradores provoca no fluxo de ar. Essas alterações têm naturalmente impacto na produção dos aerogeradores por onde o ar passa de seguida. Assim, para além da localização geográfica do parque, a localização dos aerogeradores dentro de um parque afeta a produção de energia de um parque.

A otimização de parques eólicos pode permitir o aumento da produção de energia de um parque. Para tal são usadas técnicas que permitem descobrir a localização ótima das turbinas de modo a maximizar a energia produzida.

O objectivo deste projecto é o desenvolvimento de um modelo para optimizar o layout de um parque eólico baseado no impacto esperado da localização dos aerogeradores na produção. Para isso são usadas técnicas de data mining e machine learning para obter um modelo que predirá a produção de um aerogerador baseado na sua localização, bem como outras variaveis externas.

Acknowledgements

I would like to thank my supervisors for helping me in every aspect of this dissertation. To my friends for helping me trough this very fast five years , on those exhausting nights and sometimes even days and their support in the most difficult of times. I would like to thank my parents for always supporting me even when times get rough and for providing me with a good education. Their patience, kindness and sacrifice is a notable example to everyone. To all people that I'm forgetting, sorry. You, too were an important part of these 5 years.

João da Silva Carvalho

"If you torture the data long enough it will eventually confess."

Ronald Harry Coase

Contents

1 Introduction **1**

1.1 Motivation and Objectives 1

1.2 Dissertation Structure 3

2 State of the art **5**

2.1 Wind power prediction 5

2.1.1 Physical Models 5

2.1.2 Statistical Models 6

2.2 Wake Effect 6

2.3 Data Mining 8

2.3.1 Process 8

2.3.2 Common Tasks 9

2.4 Regression 10

2.4.1 Model evaluation 10

2.4.2 Performance Estimation 10

2.4.3 Regression Trees 11

2.4.4 Support Vector Machines 12

2.4.5 Multiple Linear modeling 12

2.4.6 Random Forests 13

3 Case Study **15**

3.1 Data collection 15

3.2 Data Preparation 17

3.3 Experimental Setup 19

3.4 Additional Results 21

4 Results and Discussion **23**

4.1 Random Forest 23

4.2 Linear Regression 24

4.3 Support Vector Machines 26

4.4 Recursive Partitioning Tree 26

4.5 Discussion of Results 29

5 Final remarks and Future Work **33**

5.1 Difficulties 33

5.2 Future Work 33

References **35**

CONTENTS

List of Figures

2.1 Wake effect turbulence on wind turbines 7
2.2 Steps that compose the KDD process [9] 9
2.3 Example of a regression tree 12
2.4 13

3.1 Western Wind Resources Dataset 16
3.2 Layout of the selected turbines 16
3.3 Monte Carlo process diagram in `performanceEstimation` provided by Luís Torgo 20

4.1 Variable importance 24
4.2 Results of the simple linear regression model 25
4.3 Plots for linear modeling analysis 27
4.4 Regression Tree of the obtained model 28
4.5 Variable Importance in the Regression Tree 28
4.6 Daily NMSE of the methods used 30
4.7 Results of the analysed models 31

LIST OF FIGURES

List of Tables

4.1 Random Forest results 23
4.2 Variable Importance 24
4.3 Linear modeling Results 25
4.4 Support Vector Machine results 26
4.5 Regression Tree results 26

Abbreviations

NWP	Numerical Weather Prediction
KDD	Knowledge Discovery in Databases
NMSE	Normalized Mean Squared Errors
MSE	Mean Squared Errors
PLS	Partial Least Squares
MLP	Multi Layer Perceptron
NREL	National Renewable Energy Laboratory
CSV	Comma Separated Values

Chapter 1

Introduction

Wind energy is the transformation of the wind power to useful energy, like in windmills for the production of electrical and mechanical energy, or sails to propel ships. It is a clean energy, it can be produced in almost every region and it requires less area than most of the energetic options. [17]

Wind energy is in constant expansion and it is one of the main sources of renewable energy. Its growth motivates the energy industry and creates jobs, having, between 2007 and 2010, created 30 per cent more jobs in the sector [8]. It is estimated that in 2012, in the European Union, wind energy employs 249 thousand people, and it is expected to employ 520 thousand in 2020. For example Denmark intends to reach 50 per cent growth in 2025 [10].

A wind farm is a group of wind turbines placed in the same space. Each group can reach a few hundred turbines. The surrounding area of the farm can be used for agriculture or other purposes. A wind farm can also be placed off-shore. Comparing wind energy with the traditional ways of obtaining energy, the environmental impact is relatively smaller. This type of energy does not release any gas to the atmosphere and the energy spent for construction and manufacturing of the farm is equal to the energy that the farm produces in just 3 months, despite having a lifetime of 20 to 25 years. It is possible to optimize a wind farm by placing the turbines in places that won't interfere with the other turbines [6]. This is done by analysing the wake effect, a effect that alters the wind, making possible the optimization of the wind farm .

1.1 Motivation and Objectives

Wind behaviour is altered when passing by solid bodies, making nearby turbines less effective. This problem affects the energy production of a wind farm, making them less profitable. By placing turbines in adequate places this effect can be reduced making the turbines more efficient. This could be possible by producing a model that would predict the power generated by a turbine taking in account its location as well as the location of the surrounding turbines.

Introduction

The goal of this project is to create a model, using data mining and regression techniques, that predict the values of power production given some variables, as the position of turbines. If successful these models can be used as a base to develop a decision support system to optimize the layout of a wind farm.

1.2 Dissertation Structure

This dissertation is divided into five chapters. Inside this chapter some of the introductory concepts are explained as well as the motivation of this project.

The second chapter describes the methods that were used to develop the project. Techniques of wind power prediction and data mining are explained, as well as the concept of Wake Effect and regression methods used to develop different models.

The third chapter approaches the case study and describes the process of retrieving data and its preparation, as well as the setup of the experimental environment. Chapter 4 will present the results of the project and its discussion. The results of the various regression techniques used are presented as well as a more detailed analysis of each technique. The last chapter contains the conclusions of the project. In this chapter the difficulties felt along the project are described and the possible future work is explained.

Chapter 2

State of the art

This chapter contains the latest research made by each of the techniques used in this project. Wind power prediction models are explained to further understand how to predict wind power. The concept of wake effect is explained as it is a very important part to consider in the making of this project. The data mining process is described as it is one of the main steps of the projects development. Finally the regression techniques used to acquire models that predict the turbines power production are described to further understand its functioning.

2.1 Wind power prediction

Atmospheric conditions perform an important role in most renewable energies, being this role more relevant for wind energy. [13] Prediction models can be divided into two groups: Historical analysis of wind series and Numerical Wind Prediction Model (NWP). The approaches used are typically characterized in three groups, which will be discussed next:

- **Physical model** – Physical models try to use physical variables as much as possible to achieve the best wind speed estimation before using Model Output Statistics (weather forecasting statistical technique).
- **Statistical model** – Statistical models try to find relationship between explanatory variables, including NWP and measured data, usually employing recursive methods.
- **Combined model** – The objective is to benefit from the advantages of physical and statistical models to obtain a globally optimal performance for the examination horizon.

2.1.1 Physical Models

Several physical models based in the use of time data, have been developed to predict the speed and the power of the wind. Physical models generally use global databases of meteorologic measures or atmospheric mesoscale, but they require large computational systems to achieve good result. In the physical approach a detailed description of the low atmosphere is used to estimate the power output of the wind.

2.1.2 Statistical Models

Classical statistical forecast projections for few days are not used, as current dynamical NWP models are more accurate. There are two types of classical statistical are used to improve NWP models [12]:

- **Perfect Prog** – It uses predictions of a NWP model for future states of the atmosphere, assuming that they are perfect.
- **Model Output Statistics (MOS)** – They can include directly in the regression the influences of characteristics of different NWP models in different projections in the future. To obtain a prediction equation it is necessary to develop a set of data with historical records and records of the predictions of the NWP model.

Both use large regression equations. The advantages of Perfect Prog are the large amount of samples used as they use historical climate data, the fact that the equations are developed without the information of the NWP, being that changes in the NWP model don't change regression equations. With the improvement of the NWP models the predictions of weather will improve and the same regressions equations might be used in any NWP model. The disadvantages are that the potential predictive models should be well predicted by the NWP mode. On the other hand the advantages of the MOS are that systematic errors on the NWP are taken into account and different equations require different projection time. On the negative side it requires several years of prediction register of the NWP model and the models suffer regular changes.

2.2 Wake Effect

A wake is the region of recirculating flow immediately behind a solid object, the surrounding flux of fluid. The impact of the wake effect in the turbines leads to lower productions. There are several models that describe wake effect on wind turbines:

- **Kinematic Models**
- **Field Models**
- **Added turbulence Models**

Kinematic Models use only the impetus equation to model the deficit of the air behind a turbine. They also do not cover the change of intensity of the air turbulence behind a turbine, and sometimes a turbulence model must be used if the values of intensity are desired. Field models calculate the complete field flux through the wind farm or a part of it, if the park is regular, solving equations with a turbulence model. There are two types of field models: two dimensional and three dimensional models. As stated earlier kinematic models are combined with added turbulence models, when used for load calculations [18]. As 2.1 suggest the area immediately behind the turbine (Core Area) will maintain a costant wind speed with diminished speed. The Peripheral

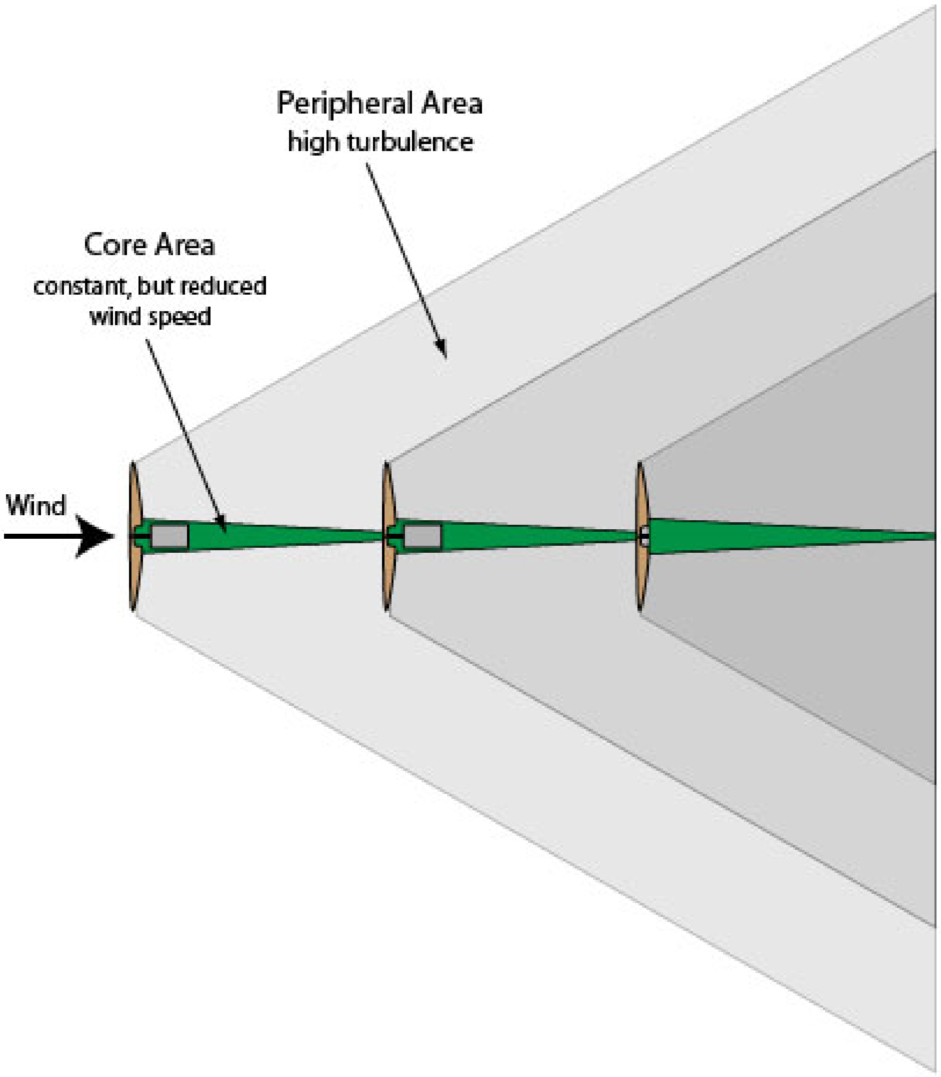


Figure 2.1: Wake effect turbulence on wind turbines

Area will have more turbulence as more turbines come with contact with the wind. As this figures shows the more solid objects that contact the wind, the more turbulence is generated, making the following turbines less effective.

2.3 Data Mining

Data Mining is the process of discovering useful knowledge like patterns, associations, changes, anomalies and significant structures from large amounts of data using artificial intelligence, machine learning, statistics and database systems techniques. Data mining is actually the core step in Knowledge Discovery in Databases (KDD) process.

2.3.1 Process

KDD is an iterative process that transforms raw data into useful information. The different steps of Knowledge Discovery in Databases are [20]:

1. Selection
2. Pre-processing
3. Transformation
4. Data mining
5. Evaluation

Selecting data is one of the most important steps. Appropriate data must be selected to perform data analysis and get useful knowledge. The data set should have enough quantity of data to perform good data mining. The pre-processing step removes noise and irrelevant data from the data set obtaining a cleaner data set. It is a very important step because it trims data and improves its quality. With the use of transformation methods, the data is prepared and transformed in appropriate form, ready for data mining, which reduce the number of effective variables selecting only useful features to optimize the goal task. The selection of the appropriate task to perform data mining is crucial, as well as the choice of the appropriate algorithm for data mining, given the problem to be solved. With this selection the data mining step is ready to be processed. The last step is the evaluation of patterns where the user interprets the mined data. If no useful pattern is found, the process might start in a previous step, making KDD an iterative process . The various steps that compose the KDD process are shown in figure 2.2.

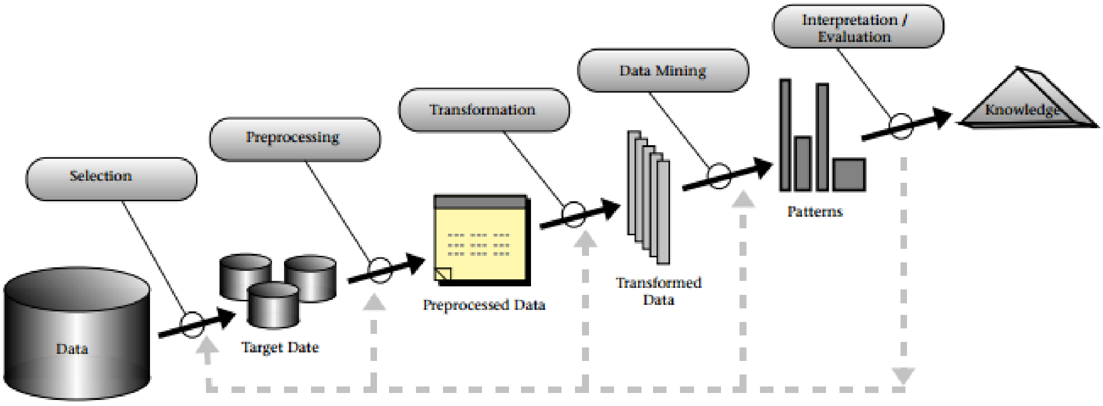


Figure 2.2: Steps that compose the KDD process [9]

2.3.2 Common Tasks

There are some common tasks that are addressed in data mining such as [9, 23]:

- Anomaly detection – Refers to the problem of finding patterns in data that do not conform to expected behavior, like in fault or fraud detection or even system health monitoring.

- Association Rule learning – It is used to discover interesting relations between variables. Its intent is to identify strong rules discovered in databases, using different measures of interestingness. A famous story about association rule mining is the "beer and diaper" story in which a survey of behavior of supermarket shoppers discovered that customers who buy diapers tend also to buy beer.

- Clustering – Clustering is a division of data into groups of similar objects. Each group, called a cluster, consists of objects that are similar to one another and dissimilar to objects of other groups [2]. For example, in the analysis of social networks, clustering may be used to recognize communities within large groups of people.

- Classification – It is a learning function that assigns a record to a predefined class. Classification tasks can be useful for example in predicting tumor cells as benign or malignant.

- Regression – Attempts to find a function which fits the data, according to an expression, into a model.

- Summarization – Its goal is to provide a more compact representation of the data set, including visualization and report generation.

2.4 Regression

The goal of regression is to model the relationship between variables by fitting a equation to the observed data. A regression model involves unknown parametersβ, independent variables X and one dependent variable Y. A regression model relates Y to a function of X and β [21].

$$Y_i \approx f(X_i, \beta) \tag{2.1}$$

To carry out regression analysis, the form of the function f must be specified. Sometimes the form of this function is based on knowledge about the relationship between Y and X that does not rely on the data.

2.4.1 Model evaluation

There are many techniques to find the accuracy of a model. The most common is the Mean Squared Error (MSE). The MSE is the quadratic difference between the predicted and the real values [7]:

$$MSE = \frac{1}{N} \sum_{i=1}^{n} (\widehat{y}_i - y_i)^2 \tag{2.2}$$

Normalized Mean Squared Error (NMSE) is also used. NMSE is an estimator of the overall deviations between predicted and measured values and it is defined as:

$$NMSE = \frac{1}{N} \sum_i \frac{(P_i - M_i)^2}{\overline{PM}} \tag{2.3}$$

$$\overline{P} = \frac{1}{N} \sum_i P_i \tag{2.4}$$

$$\overline{M} = \frac{1}{N} \sum_i M_i \tag{2.5}$$

NMSE can show very well the difference among models. If a model has very low NMSE than it is performing well, but the opposite does not means that the model is completely wrong. The values represent the percentage of the overall deviations between the predicted and the real values [4].

2.4.2 Performance Estimation

Assessing how well data mining models perform against real data is a very important step. It is crucial to validate the data mining models by understanding their qualities and characteristics before deploying them. There are several approaches to evaluate a data mining model, such as the use of multiple statistical measures to determine if there are problems in the data or in the model, the separation of data in sets of test and training to assess the precision of the predictions or the revision of the results to determine if the patterns have significance. Evaluating model performance can't be done using the data set used for training because it can lead to overoptimistic and overfitted models (when a model is more accurate in fitting known data but less accurate in predicting new data). The use of the test sample is required only if the task to perform is a prediction [11, 15, 19].

2.4.3 Regression Trees

Building a regression tree is similar to building a classification tree, but in regression trees there is no need of assigning objects to classes. The evaluation used in a regression tree differ from those used in a classification tree, despite the similarities in the construction of the tree.

There are three components regarding the construction of a regression tree:

- A set of logical questions with binary response (yes or no).
- The choice of the split criteria to choose the best split on a variable.
- Terminal nodes with the summary statistics.

The last component is only used by the regression tree as it outputs numeric results of the dependent variable, unlike the classification tree, that outputs a class.

The main goal of the regression tree is to build a tree with the predictor or prediction rule. There are two purposes to the predictor, which are the accurate prediction of the dependent variable in the future with new independent values and find the relationship between variables.

To construct the tree the variance in terms of the dependent variable is detected in the data set, which is then purified. This is done by recursively partitioning the data until reaching the terminal nodes, where the data is more homogeneous. The terminal node has the value of the dependent variable which is then used as the prediction result [24]. In 2.3 is possible to see an example of a regression tree.

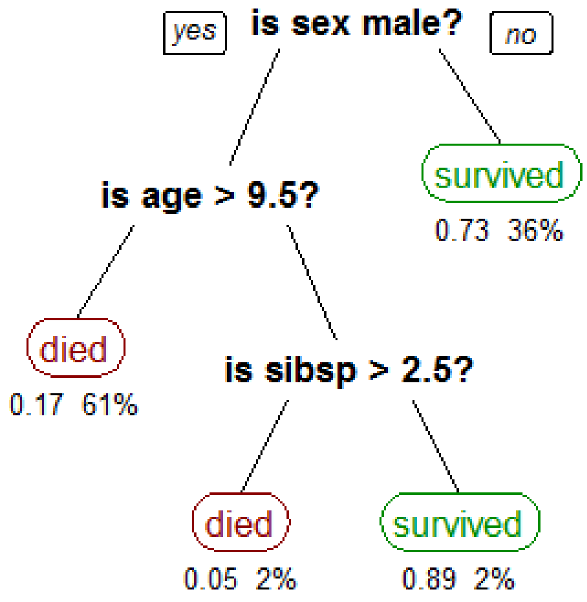


Figure 2.3: Example of a regression tree

2.4.4 Support Vector Machines

Support Vector Machines (SVM) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The training data has a set of input vectors, with each one having component features, and they are paired with a label. It is important to evaluate the goodness of the model before actually using it as this technique allows tuning of parameters. The SVM finds a hyperplane or a set of hyperplanes that maximizes the distance between classes, being that the bigger the distance the better. SVM can also perform non linear classification using what is called kernel trick, mapping their inputs in high-dimensional spaces [5].

It is possible to use SVM for regression maintaining all its features, but instead of trying to separate the objects, in regression, the SVM attempts to fit all the objects inside a defined number for error. This method can ignore error in a certain range making it having the called soft-margin. The goal is to minimize the minimum error to make the SVM more accurate [16].

In 2.4 it possible to see a transformation from a non-linear to a linear separation, being ϕ the kernel function that allows the process.

2.4.5 Multiple Linear modeling

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a dependent variable by fitting a linear equation to observed data. Every value of the

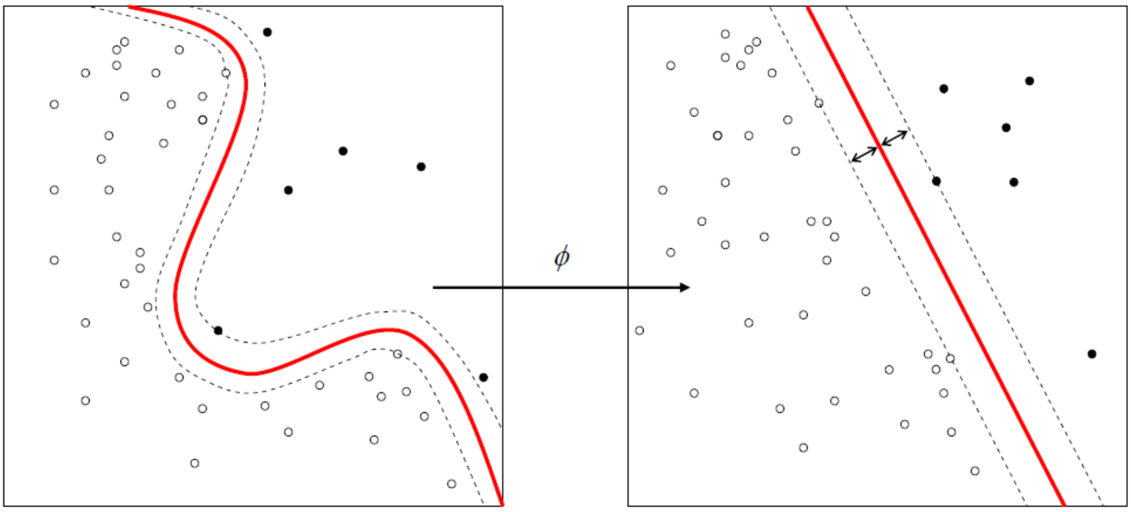


Figure 2.4

independent variable X is associated with a value of the dependent variable Y. The regression line for p explanatory variables $X_1 ... X_p$ is defined to be [1]:

$$U_Y = \beta_0 + \beta_1 X_1 + ... + \beta_p + X_p \tag{2.6}$$

This line describes how the mean response U changes with explanatory variables.The fitted values $b_0, b_1, ..., b_p$ estimate the parameters $\beta_0, \beta_1, ..., \beta_p$ of the regression line. Since the observed values for y vary about their means U_Y, the multiple regression model includes a term for this variation, the residuals. It represents the deviations of the observed values Y from their means U_Y. The notation for the model deviations is ε:

$Y_i = \beta_0 + \beta_1 X_i 1 + ... + \beta_i p X_i p + \varepsilon_i$ for $i = 1, 2, ..., n$

2.4.6 Random Forests

The random forest is an ensemble approach for classification and regression. Ensembles are approaches that combine the decisions of multiple models into a single decision. The random forest starts with a standard machine learning technique called a decision tree which, in ensemble terms, corresponds to our weak learner. A weak learner is defined to be a classifier which is only slightly correlated with the true classification. In a decision tree, an input is entered at the top and as it traverses down the tree the data gets bucketed into smaller and smaller sets. The process of training a Random Forest system is:

1. Obtain a subset with N cases at random. It should be about 66% of the total set.
2. At each node:
   (a) For number m predictor variables are selected at random from all the predictor variables

(b) The predictor variable that provides the best split, according to some objective function, is used to do a binary split on that node.

(c) At the next node, choose another m variables at random from all predictor variables and do the same.

Depending upon the value of m, there are three slightly different systems:

1. Random splitter when $m = 1$
2. Breiman's bagger where m = total number of predictor variables
3. Random forest when $m <<$ number of predictor variables

While running, when a new input is entered into the system, it is run down all of the trees. The result may either be an average or a weighted average of all of the terminal nodes that are reached.

Random forest runtimes are quite fast, and they are able to deal with unbalanced and missing data. Random Forest weaknesses are that when used for regression they cannot predict beyond the range in the training data, and that they may overfit data sets that are particularly noisy [3].

Chapter 3

Case Study

In this chapter the process of data collection is explained. The sources and the modeling of data are explained as well as the selection of data and the reasons of the selection. The data preparation step is explained in detail as this is a crucial step in the making of the project. To finish the experimental setup is described. This approaches the variables used as well as the steps in the preparation of the running of he experiments.

3.1 Data collection

In any data analysis the most important thing is the data. Thus data collection is a very important step. The data set must be carefully selected to achieve good results.

In this case, to analyse the impact of the wake effect on the production of wind energy, an ordered time series of data related to wind turbines was required. The data sets used in this case were retrieved from the Western Wind Resources Data set developed by United States National Renewable Energy Laboratory(NREL), which gives access to more than 30,000 sites. The data sets are displayed in a map representing the United States and each turbine icon represents a part of a site consisting of ten turbines. The data is modeled by 3TIER, a renewable energy assessment and forecasting enterprise, using the Weather Research and Forecasting model to downscale the data, allowing one turbine to represent ten. The groups of turbines are two to five kilometers apart and the wake effect only fades from twenty kilometers and beyond [14]. From now on the group of ten turbines will be designated as turbine. It is possible to see a short area of the map in figure 3.1

To fully analyse the impact of Wake effect , a small set of turbines was selected. All the chosen data sets refer to the year 2006, and the records are 10 minutes apart, making the total number of 52560 records. Eight turbines were selected from the South Park, National Park and their relative placement to each other was taken into account for the analysis. The display of the turbines can be seen in figure 3.2. This allows the two central turbines to be analysed using both their own data set and data sets of from neighbour turbines.

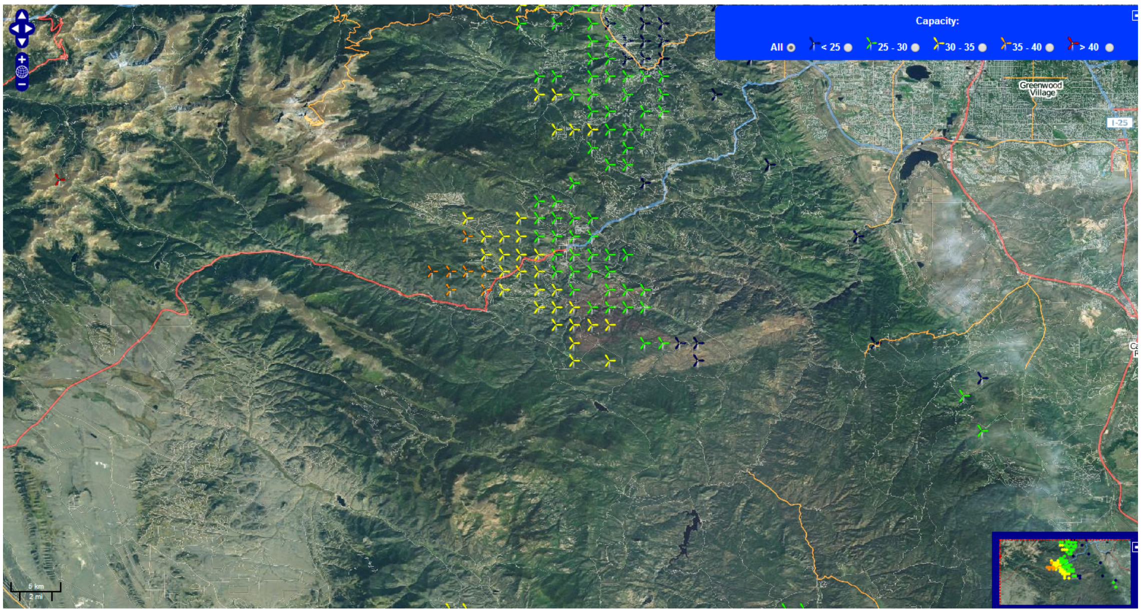


Figure 3.1: Western Wind Resources Dataset

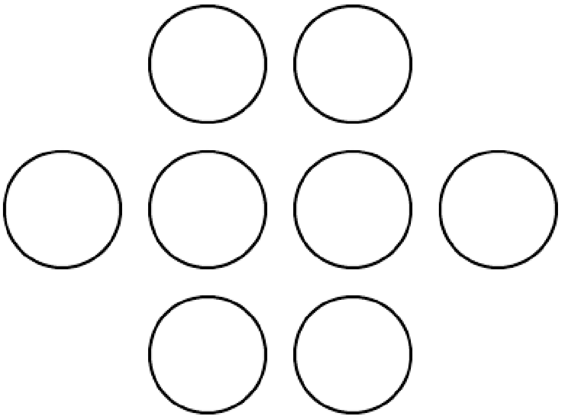


Figure 3.2: Layout of the selected turbines

In order to perform this analysis, weather records are needed. Variables like the wind direction and the wind speed are crucial to understand how relevant is the wake effect in wind power production. This data is also obtained using NREL, which offers the average wind speed and the average wind direction with a five minutes interval. The meteorologic data and the data retrieved from the turbines are not enough to perform the required analysis as the display of the turbines plays an important role in the regression tasks because its position is directly related to the wake effect. Depending on the winds direction the values of the production should differ, making this display allied with the direction of the wind one of the main focus of the project.

3.2 Data Preparation

After data is retrieved it needs to be treated in order to be correctly processed, to obtain a good model for prediction of energy production to better understand how wake effect interferes with it it. The interval between meteorologic data and the power output data set was different so, to correct that, the interval of the meteorologic data was extended to 10 minutes, erasing even rows of the initial data set. The data sets were retrieved as CSV files (Comma Separated Values), and using R, were stored in three tables in a Postgresql relational database:

- **Values** – Table with the power output of each turbine.
- **Meteodata** – Table containing the South Park average wind direction and average wind speed values.
- **Neighbours** – Contains the ID of the neighbour turbines as well as the latitude and longitude of the respective turbine.

There are several options for data base creation like MySQL, SQLite, etc but in this dissertation Postgresql is chosen due to past experience and its ease in creating and managing databases. The use of R is suggested by the supervisors as it is a powerful tool in data mining and the use of several packages that would facilitate the data analysis. The neighbour turbines are considered to be the turbines that are North, South, East and West from the target turbine. Even though all the variables were stored, because they might be useful, in the Values table only the SCORE-lite from each turbine and neighbour turbines were used. The SCORE process uses observed deviations from a mean value to create probability density functions of deviation from some central point. SCORE is run for each individual turbine and produces a time series of data for each turbine. The individual turbine time series are then aggregated to represent sub-project groupings or are summed up to model the entire project output. However, to model the output for more than 30,000 individual turbines is extremely time consuming. To solve this problem SCORE-lite was developed. SCORE-lite uses the “rated” power output, calculated by converting wind speed to power output through a simple rating curve, and modify it such that the overall characteristics are more approximate those observed in reality.

The relative position of the neighbour turbines must be considered as is an important variable when creating the regression model. This variable combined with the average wind direction would represent the wake effect impact on the production as if one turbine presents lower values in production, it might be associated with the position of the surrounding turbines.

The average wind direction of the park is given in degrees from North, so each neighbour average wind direction is calculated according to it's relative position:

- **North Neighbour** – The north neighbour wind direction will be the same as the parks since it's average wind direction is given from north.
- **South Neighbour** – For the south average wind direction 190 degrees are added to the park's direction.
- **West Neighbour** – 90 degrees were added to the west neighbour to the wind direction wind values.
- **East Neighbour** – 270 degrees were added to this turbine values'.

All values were submitted to the modulus operation with 360, to obtain values within the 360 interval. To find the relative position of the neighbour the angle between the vector $\vec{w} = (1,0)$ and the vector between the central turbine and the neighbour in question $\vec{v}$. To do so $\vec{v}$ must be calculated. This vector will be retrieved subtracting the longitude and the latitude of the central turbine and the neighbour in question so $\vec{v} = ((longitude_n - longitude_c), (latitude_n - latitude_c))$. It is now possible to find the angle between the two vectors with:

$$cos\theta = \frac{v \cdot w}{||v|| ||w||} \tag{3.1}$$

With this, it is now possible to assign the resulting angle to a neighbour fitting them in an interval of angles:

- **North** – Between 315 degrees and 45 degrees.
- **South** – Between 135 degrees and 225 degrees
- **East** – Between 225 degrees and 315 degrees.
- **West** – Between 45 degrees and 135 degrees.

With this it is possible to identify the location of the neighbour. To perform the regression and obtain a model that predicts the power output correctly, the variables must contain values which are known at the time the prediction is generated.The data was prepared such that each observation contains the values of the independent variables for the data of the central turbine and their neighbours' at time t, but the value of the dependent variable, ie the production of tbe central turbine at time $t+n$ being n the prediction horizon. This will train the model to predict the power outputted correctly for t. The final data set used for the regression comprises the following variables:

- SCORE-lite of the central turbine as well as from its neighbours.
- The average wind direction of each turbine.
- The park's wind speed.
- The park's wind direction.

3.3 Experimental Setup

To learn and evaluate the regression models, using the data set that was obtained in the data preparation step, R was also used. This statistical programming language provides multiple algorithms to analyse the performance of a model as well as constructing the model referring to the regression method.

The evaluation of the models performance was made using the R package `performanceEstimation` [22]. This package provides flexible tools for performance estimation and experimental comparison of predictive models. It enables the comparison of the performance of multiple regression algorithms and the selection of the best one. To do so, a function that has the same name as the package that was used. The `performanceEstimation` function has three arguments: Predictive Task, WorkFlow and Estimation Methodology.

Predictive Task

This step will determine the variables that are used and the regression formula used in each work flow. To observe the energy production, that is represented by SCORE-lite, a formula to evaluate the model is needed. In this case that formula is *score~.*, being that using the dot is the command to choose all the other variables as independent variables.

WorkFlow

The goal is to estimate the predictive performance of a proposed work fow to solve the task, by using different samples to increase our confidence on the estimates. This workflow consists on the process of obtaining a model from a given training sample and then use it to obtain predictions for the given test set. Because the data set consists of a ordered time series of records the chosen workflow was timeseriesWF. This workflow function implements two different approaches to the problem of training a model with a set of time-dependent data and then use it to obtain predictions for a test set in the future. A time series workflow needs a time window in order to train and test data. There are two possible choices:

- **Sliding Window** – Uses data occuring in the last L time steps.
- **Growing Window** – Keeps increasing the original training window with the newly available data points. The models are obtained with increasingly larger training samples.

In this case the sliding window was chosen due to its lower time and computation costs. To execute the workflow an algorithm must be chosen. This function is called with a formula in the first argument and the training set in the second. The selected algorithms are referent to the chosen methods to perform and evaluate the regression models.

Evaluation of models are done using a evaluator. There are several evaluators, but in this case the Normalised Mean Square Error (NMSE) was used as it gives a overall overview of the difference between the predicted and the real values.

Estimation Methodology

There are different ways of providing reliable estimates of the predictive performance of a model. The `performanceEstimation` package implements some of the most common estimation methods. In this case Monte Carlo experiments were used because the original order of the observations is respected and train and test splits are obtained such that the testing samples appear "after" the training samples, thus being the methodology of choice when comparing time series forecasting models. Additionally, multiple samples are used to obtain a more reliable estimate of the performance. Monte Carlo experiments operate as follows:

1. Given a data set spanning from time t_1 till time t_N.
2. Given a training set time interval size L and a test set time interval size F, such that $T+F<N$
3. Monte Carlos experiments sample R randomly time points from the interval $[t_{1+T}, t_{N-F}]$
4. For each of these R time points they generate a training set with data in the interval $[t_{R-T+1}, t_R]$, and a test set with interval of $[t_{R+1}, t_{R+F}]$

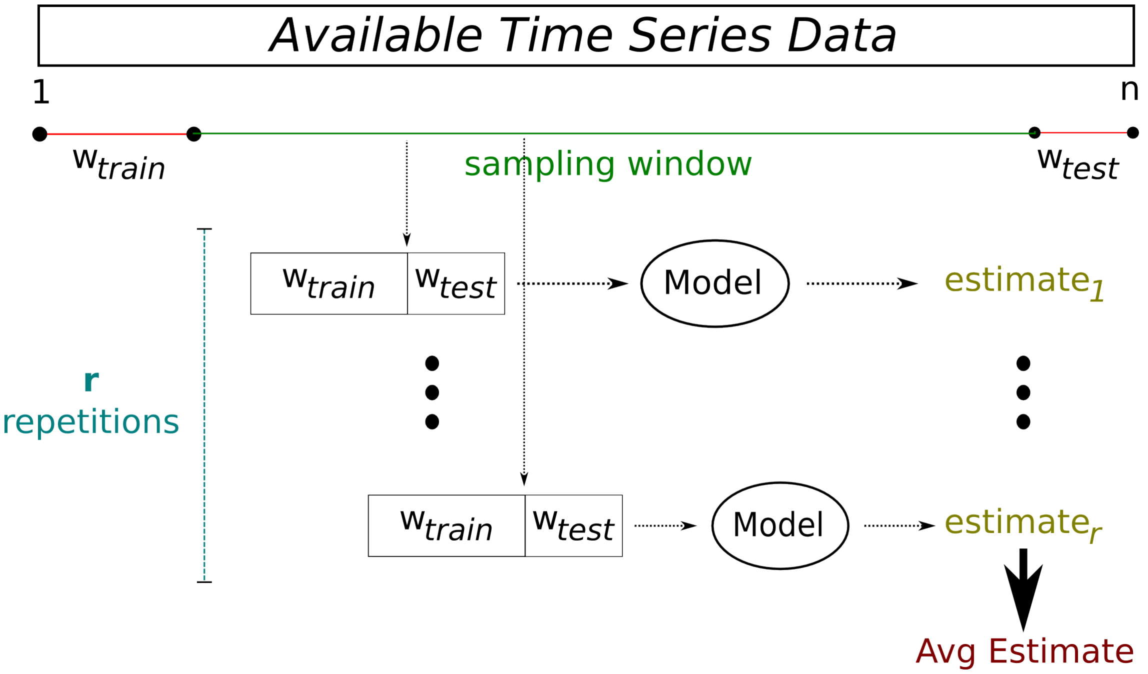


Figure 3.3: Monte Carlo process diagram in `performanceEstimation` provided by Luís Torgo

Using this process R train+test cycles are carried out using the user-supplied workflow function, and the experiment estimates result from the average of the R scores as usual. Monte Carlo Experiments are used by including as third argument McSettings(), in which the arguments `nReps`, `szTrain` and `szTest` were used. All the other arguments are given by default.

The nReps argument represents the number of repetitions of the Monte Carlo experiment while szTrain and szTest represent the percentage of cases used in the training samples and the percentage used in test samples respectively. It is possible to understand this better by looking at 3.3). The values used in this case study are ten for nReps, 50 per cent for training size and 25 per cent for testing. The obtained values are then observed and conclusions are taken from them.

3.4 Additional Results

After this experiment a second set of experiments was made. In these experiments the data set was divided into 19 data sets, each data set representing a day, and for each data set the chosen regression methods were applied. As opposed to the first experiment where only four data sets were analysed, in this experiment the four regression methods are applied to 19 data sets. This is useful for comparing the overall results with short-term results as methods might perform better with less data, or even to observe discrepancies in the results.

Case Study

Chapter 4

Results and Discussion

In this chapter, we will present and discuss the results from the methods used for regression. The analysis of each method is done and the results explained.

First Experiment Results

4.1 Random Forest

After performing an analysis of the performance of the Random Forest model with the given data set the following results were given:

	MSE	NMSE
average	140.017	0.918
standard	9.992	0.074
min	118.741	0.781
max	154.674	1.015
invalid	0.000	0.000

Table 4.1: Random Forest results

The average NMSE is almost 92 per cent meaning that a the predictions were very distant from the actual values. The importance of the variables when splitting in the tree node can be seen in the node purity. It is possible to observe a table with the variable importance at 4.2, as the variables represents the power output, the average wind speed of a certain neighbour, which is determined by the number, as the park variables.

This is the total decrease in node impurity from splitting on the variable. The Node Impurity, the more important is the variable. In 4.1 we can see that the average wind speed of the park is the most important independent variable. This is a logical association because we can easily understand that the wind speed will have a great impact in the production of wind energy. The second most important is `power2`, which is energy production of the northern turbine. This means

Variable Name	Node Impurity
power1	42,751.85
avgwinddirection1	34,735.23
power2	51,039.00
avgwinddirection2	36,100.88
power3	37,006.78
avgwinddirection3	37,571.30
power4	42,551.40
avgwinddirection4	34,104.70
avgwindspeed	70,664.59
avgwinddirection	34,372.41

Table 4.2: Variable Importance

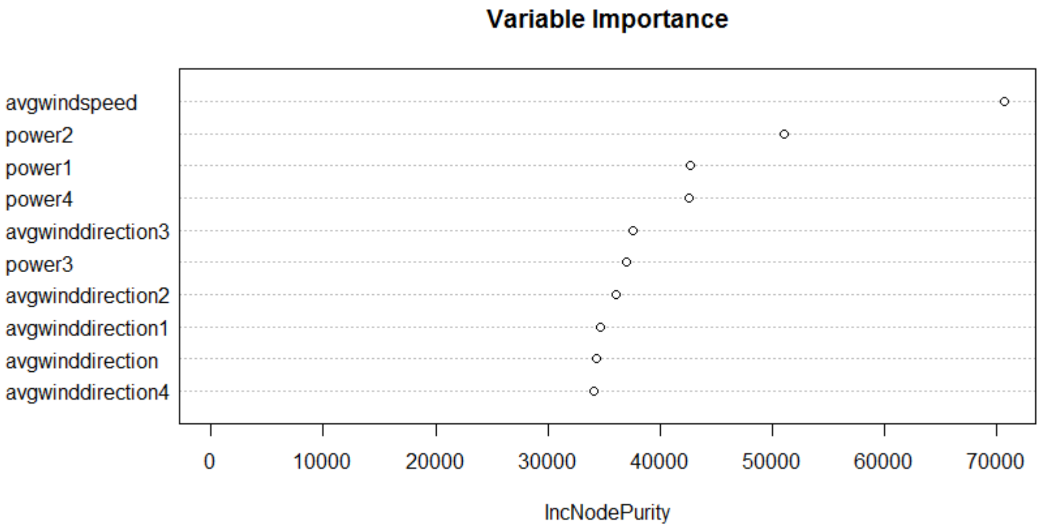


Figure 4.1: Variable importance

that when the production of the north neighbour is affected the production of the central will be, relatively to the other turbines, more significative.

4.2 Linear Regression

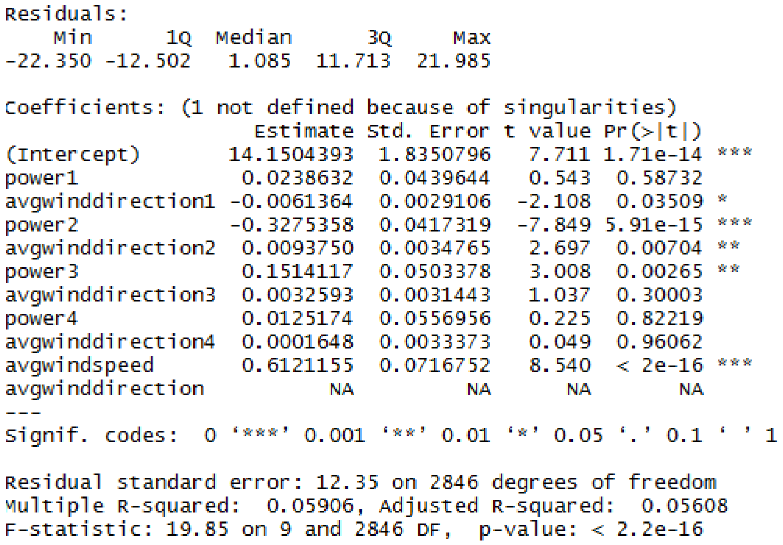
In R a linear regression is obtained using `lm` to fit the values. Executing `performanceEstimator` the following results in 4.3

The output of the regression models is the following are represented in 4.2

Analyzing the residuals can lead to some interesting things about the model. The residuals are the difference between the actual values of the target variable and the predicted values $y - \widehat{y}$. Ideally, for most regressions, the residuals should like a normal distribution when plotted. If our residuals are normally distributed, this indicates the mean of the difference between our predictions

	MSE	NMSE
average	151.933	0.995
standard	8.196	0.038
min.	141.945	0.934
max.	167.313	1.072
invalid	0.000	0.000

Table 4.3: Linear modeling Results



```
Residuals:
    Min      1Q  Median      3Q     Max 
-22.350 -12.502   1.085  11.713  21.985 

Coefficients: (1 not defined because of singularities)
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)       14.1504393  1.8350796   7.711 1.71e-14 ***
power1             0.0238632  0.0439644   0.543  0.58732    
avgwinddirection1 -0.0061364  0.0029106  -2.108  0.03509 *  
power2            -0.3275358  0.0417319  -7.849 5.91e-15 ***
avgwinddirection2  0.0093750  0.0034765   2.697  0.00704 ** 
power3             0.1514117  0.0503378   3.008  0.00265 ** 
avgwinddirection3  0.0032593  0.0031443   1.037  0.30003    
power4             0.0125174  0.0556956   0.225  0.82219    
avgwinddirection4  0.0001648  0.0033373   0.049  0.96062    
avgwindspeed       0.6121155  0.0716752   8.540  < 2e-16 ***
avgwinddirection          NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.35 on 2846 degrees of freedom
Multiple R-squared:  0.05906,	Adjusted R-squared:  0.05608 
F-statistic: 19.85 on 9 and 2846 DF,  p-value: < 2.2e-16
```

Figure 4.2: Results of the simple linear regression model

and the actual values is close to 0 (good) and that when the model misses, it misses both short and long of the actual value, and the likelihood of a miss being far from the actual value gets smaller as the distance from the actual value gets larger. The lower the residual standard error the best, given that the goal is to minimize error to obtain a better model, taking care not to overfit the model. `Pr` gives the probability that the variable is not relevant. The smaller the value of Pr the more important the variable is. In this model is possible to see that the `avgwindspeed` is the most important, followed by `power2` that represents the power production of the southern turbine. This can also be checked when looking at the asterisks in the last column. The amount of asterisks represents the importance of that variable.

In 4.3 (a) it is possible to see the Residuals vs Fitted plot. The points should be randomly scattered around the center line as having all points on one side would violate linearity. With this we can infer that the model is better at fitting bigger values, because the red line is, at first, deviating from the center line, stabilizing on bigger values. The Normal Q-Q plot 4.3 (b) helps to analyze whether the distribution of residual error is normal. The points should be along the the diagonal line or else it means that the residual error is not uniformly distributed. In this case the points are not along the line meaning that the model does not have a uniform distribution of the residuals.

4.3 Support Vector Machines

In R the use of a common Support Vector Machines technique is done using the `svm` function.

	MSE	NMSE
average	194.136	1.272
standard	11.622	0.093
min	171.509	1.128
max	216.027	1.449
invalid	0.000	0.000

Table 4.4: Support Vector Machine results

These results were given with a SVM with a radial kernel and 2646 support vectors were generated. Moreover, it must be pointed out that differences on peaks have a higher weight on NMSE than differences on other values. Parameter tuning is very important for SVM and we had no time to do it. This possibly explains the low quality of the results obtained

4.4 Recursive Partitioning Tree

Using the `rpart` function in R will result in a regression model using a simple Recursive Partitioning Tree. The results for this technique are presented in Table 4.5:

	MSE	NMSE
average	161.004	1.055
standard	9.859	0.061
min	144.638	0.951
max	179.413	1.154
invalid	0.000	0.000

Table 4.5: Regression Tree results

The average predicted values greatly differ from the actual values, rendering high NMSE values. The regression tree referring to this model, 4.4, displays the classification, the probability of the class conditioned on the node and the percentage of observations used at that node. With the analysis of this tree is possible to understand that when the power generated by the southern turbine is 3.1 MW or more (representing 71 per cent of the cases) we have two situations: when the average wind speed is less than 11, obtaining lower values for energy production; This represents 62 per cent of the cases. Otherwise the generated power takes higher values. This is logical as wind speed is related with the production of energy.

If the power production in the southern turbine is less than 3.1 MW higher values for production. The eastern turbine production will be related to the production of the central turbine as further analysis of the data revealed that the wind as in its direction.

Figure 4.5 supports the regression model as the most important variables are the ones that are used in the regression tree.

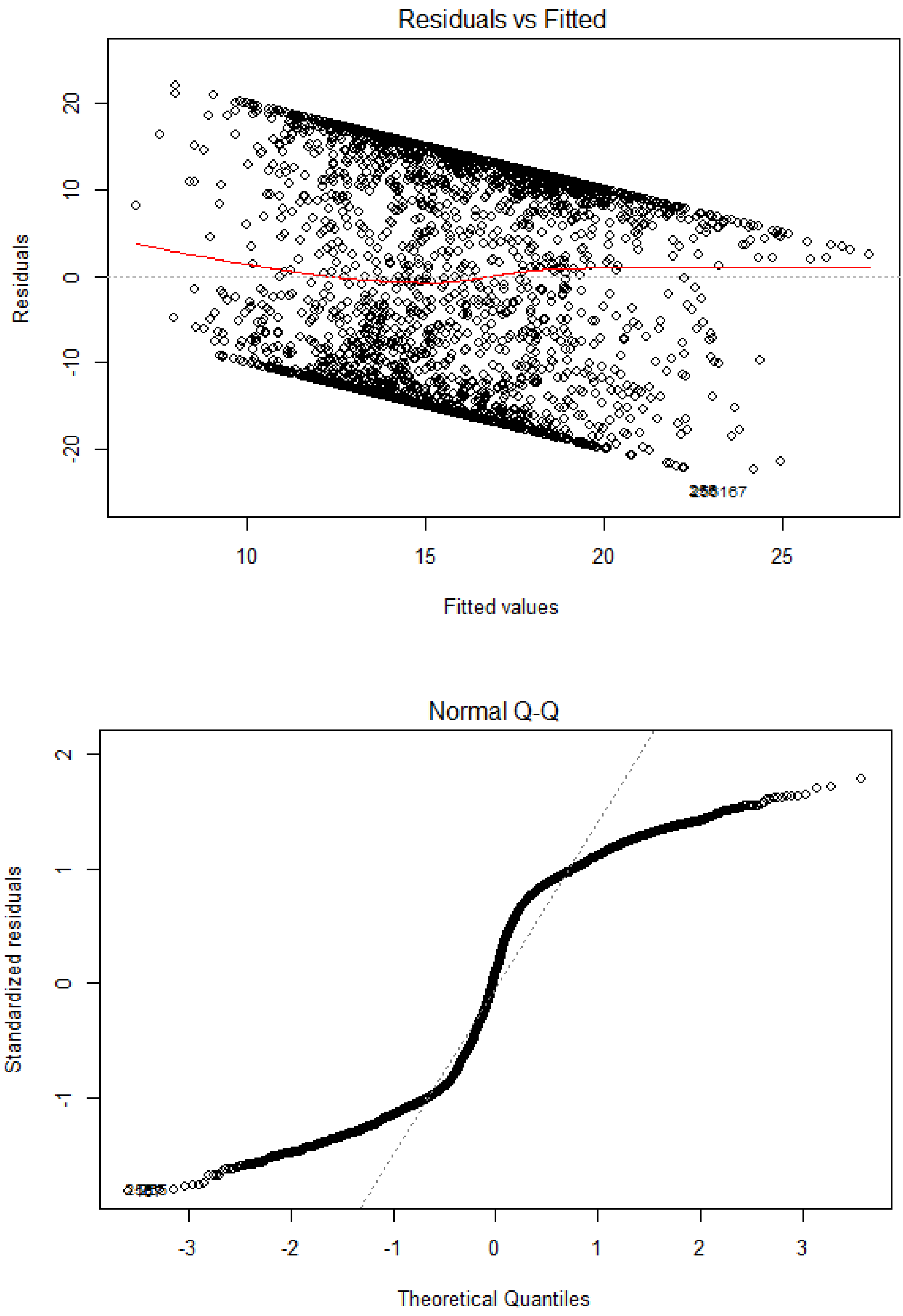


Figure 4.3: Plots for linear modeling analysis

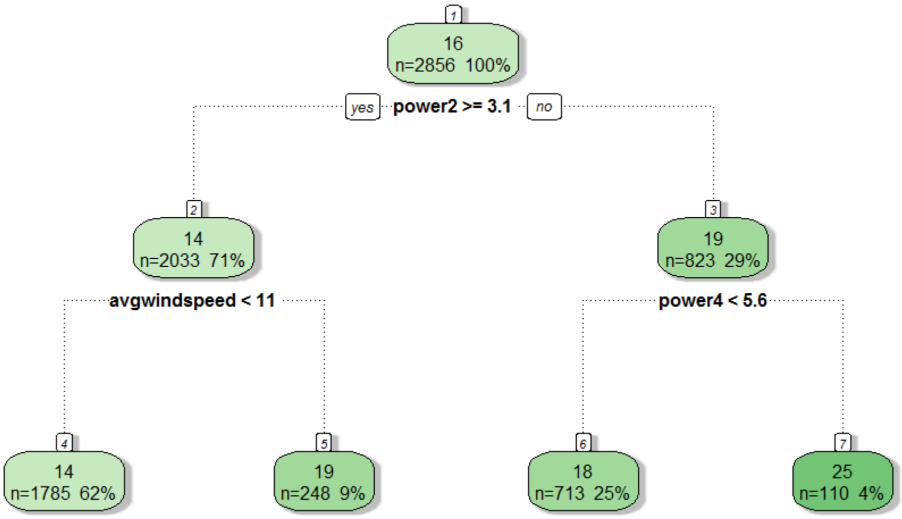


Figure 4.4: Regression Tree of the obtained model

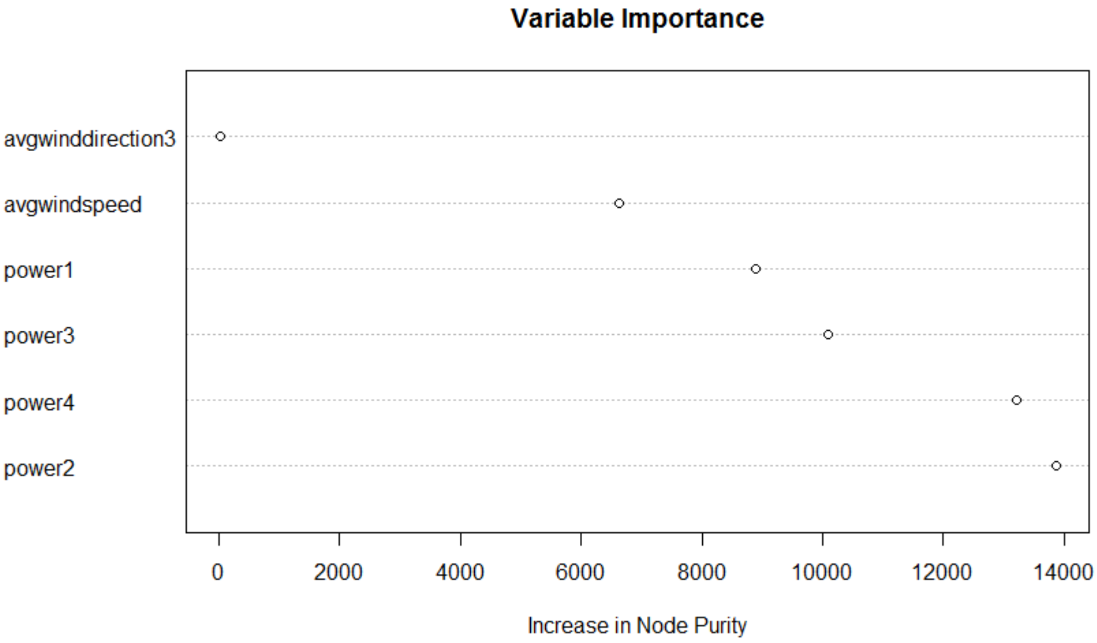


Figure 4.5: Variable Importance in the Regression Tree

Second Experiment Results

With this experiment interesting results were obtained. Despite the Linear Modeling and the Regression Tree method presenting similar results to the overall set, it is possible to see on figure 4.6 that the Random Forests method is not the best at predicting daily productions. It is easy to see that the method with the best results is the SVM. This might have to do with the fact that the SVM didn't suffered any tuning and could produce better results with it.

4.5 Discussion of Results

First Experiment

The results obtained by all the algorithms are, on average not good. The NMSE values are large, meaning that a large amount of data is predicted incorrectly making the regression model not reliable. We can see in 4.7 that the model with the lowest NMSE is Random Forests. This means that Random Forests produced, on average, more accurate results than the other techniques.

Second Experiment

To further analyse all methods the data set was divided in days and the values for daily NMSEs was calculated. With this is possible to analyse more closely the variance of the NMSE throughout the nineteen days. In 4.6 is possible to see that SVM is the method with the lowest values for daily NMSEs. Comparing to overall results, where the SVM is the method with the worst results, it is possible to conclude that SVM is better for short-term predictions (less records) . Relatively to the other methods, Random Forest continues to be the one with the least NMSE and the chosen method to model the data set. It is possible to observe high peaks on 4.6 in all methods. This can be related to a abrupt variance in the meteorologic enviornment of the park, rendering incorrect predictions.

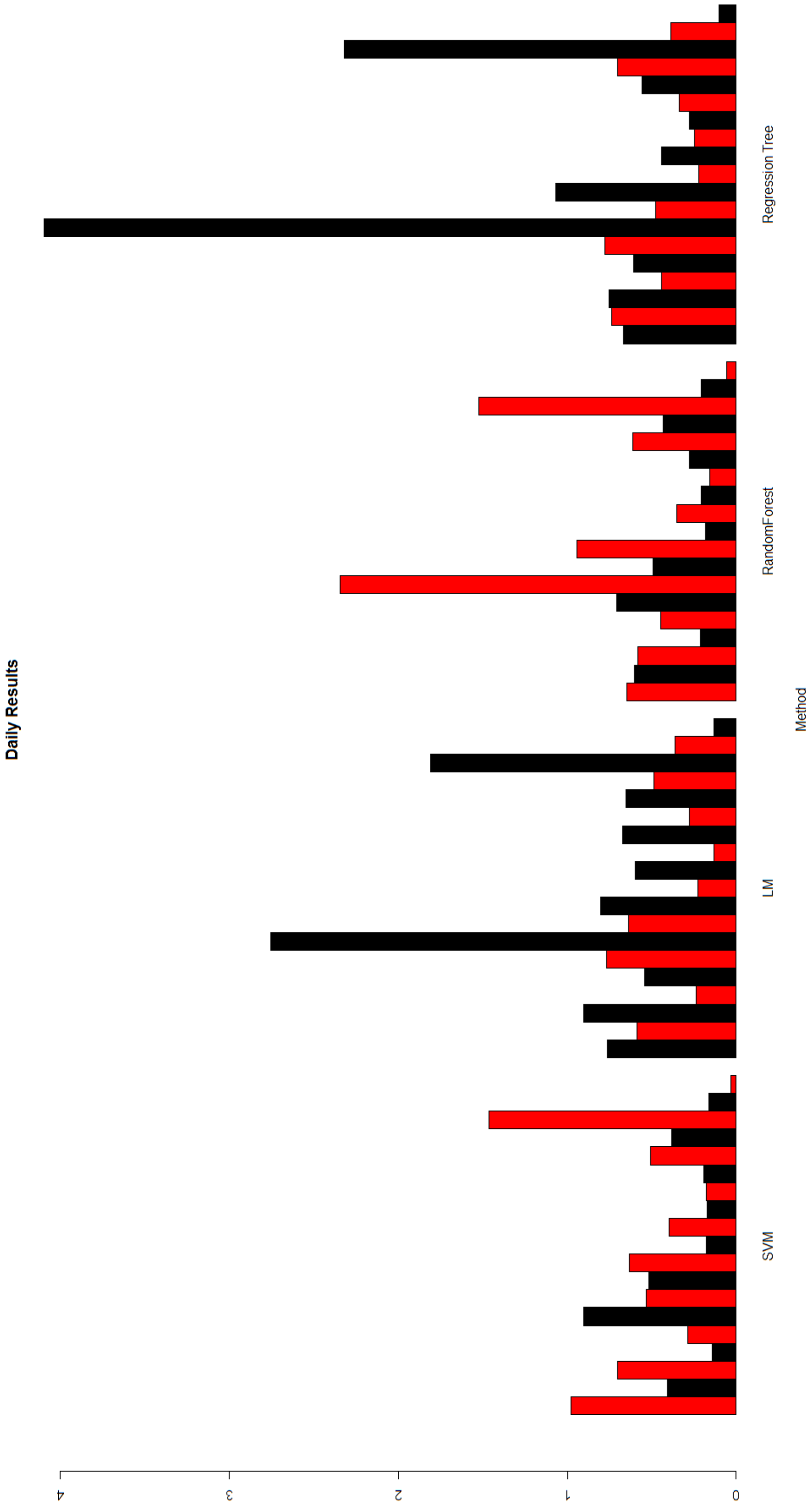


Figure 4.6: Daily NMSE of the methods used

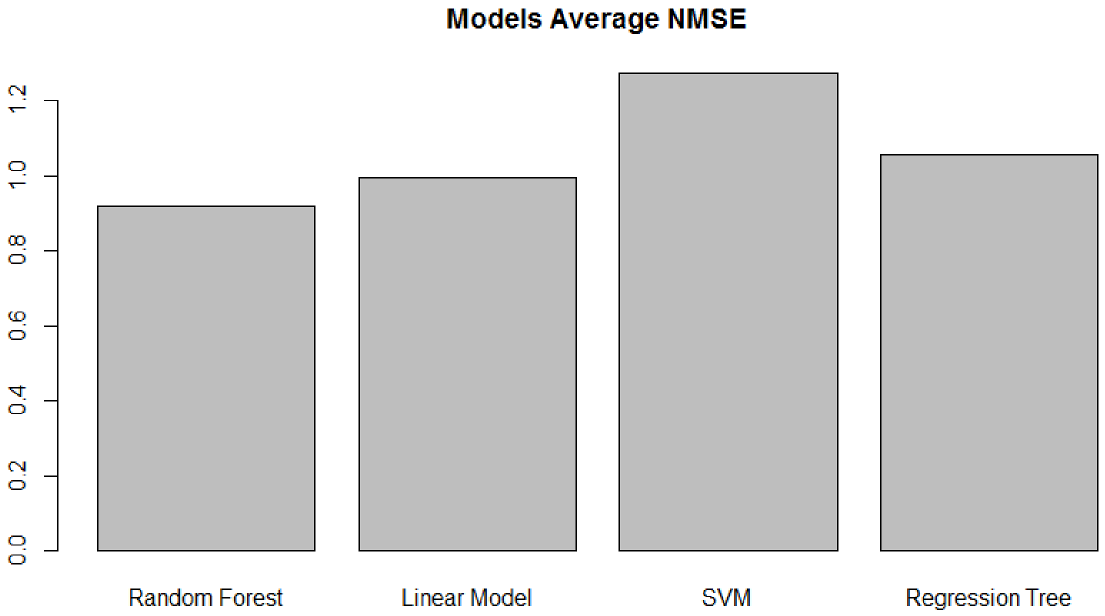


Figure 4.7: Results of the analysed models

Chapter 5

Final remarks and Future Work

Wind energy is one of the most prominent renewable energies in the current days. Understanding how the Wake Effect affects the efficiency of a wind farm may lead to a more profitable parks. This project builds a model that relates the production of wind energy in a turbine to some independent variables, with which is it would be possible to predict the production of a turbine based on its location. The final results were not sufficiently good to employ the models in practice.

5.1 Difficulties

It was not possible to collect data concerning the production of individual turbines. The output of the regression results could be better if data of a single turbine instead of a data retrieved from a model. A off-shore data set would be ideal, because uncontrolled variables like irregular terrain would be greatly reduced. This kind of data sets were not available for free.

The construction of the data set was a step that took an unexpected amount of time, because its format was not indicated for the data mining task. Data preparation was also complex. For instance obtaining the location of the turbine was a challenge because obtaining it from a group of turbines didn't supplied correct information about the location of each turbine in the group.

5.2 Future Work

With a model that correctly produces accurate predictions of the energy production of a wind turbine, it possible to improve the efficiency of a wind farm. With the use of new variables like the individual production of a turbine instead of modeled values or windspeed of a offshore park it would be possible to obtain better results. Tuning parameters in some methos is crucial to obtain a model that would predict better values, making the model more reliable. With the use of artificial intelligence techniques it is possible to obtain a layout for the wind farm that minimizes the effect that the turbulence caused by the wake effect. This would lead to the creation of a decision support system that would help constructors to optimize the location of turbines in order to achieve more profitable results.

References

[1] Multiple linear regression. Available at http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm.

[2] Pavel Berkhin. A survey of clustering data mining techniques.

[3] Arshavir Blackwell. A gentle introduction to random forests, ensembles, and performance metrics in a commercial system.

[4] European Comission Joint Research Center. Available at http://rem.jrc.ec.europa.eu/RemWeb/atmes2/20b.htm.

[5] Yiming Ying Colin Campbell. *Learning with Support Vector Machines*. Morgan and Claypool, 2011.

[6] Dr Mark Diesendorf. Why australia needs wind power. *Dissent*, pages 43–48, 2003.

[7] Geoge C. Runger Douglas C. Montgomery. *Applied Statistics and Probability for Engineers 6th edition*. Wiley, 2014.

[8] EWEA European Wind Energy Association. Wind energy facts. Available at http://www.ewea.org/wind-energy-basics/facts/.

[9] Gregory; Smyth Padhraic Fayyad, Usama; Piatetsky-Shapiro. Data mining to knowledge discovery in databases. *AI Magazine*, 1996.

[10] GWEC Global Wind Energy Council. Disponível em http://www.gwec.net/.

[11] Douglas M. Hawkins. The problem of overfitting. Technical report, School of Statistics, University of Minnesota, 2003.

[12] Daria Kluve. Statistical weather forecasting, an indepent study, from statistical methods in the atmospheric sciences by daniel wilks.

[13] Earth System Research Laboratory. Available at http://www.esrl.noaa.gov/research/renewable_energy/.

[14] Charlotte B. Hasager Merete Bruun Christiansen. Wake effects of large offshore wind farms identified from satellite sar, 2005.

[15] Microsoft. Testing and validation (data mining). Available at http://technet.microsoft.com/en-us/library/ms174493.aspx.

[16] John Shawe-Taylor Nello Cristianini. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Press Syndicate of the University of Cambridge, 2000.

[17] Department of Environment New South Wales Government. The wind energy fact sheet, November 2010.

[18] Douwe J. Renkema. Validation of wind turbine wake models, using wind farm data and wind tunnel measurements. Master's thesis, Delft School of Technology, 2007.

[19] Saed Sayad. Model evaluation. Available at http://www.saedsayad.com/model_evaluation.htm.

[20] SQLDataMining. Steps of the knowledge discovery in databases process. Available at http://www.sqldatamining.com/index.php/data-warehousing/steps-of-the-knowledge-discovery-in-databases-process.

[21] Alan O. Sykes. An introduction to regression analysis. The Inaugural Coase Lecture.

[22] Luís Torgo. Available at http://cran.r-project.org/web/packages/performanceEstimation/.

[23] Arindam Banerjee Varun Chandola and Vipin Kumar. Anomaly detection: A survey. Technical report, Department of Computer Science and Engineering,University of Minnesota, 2007.

[24] Patrick Webb Yisehac Yohannes. *Classification and Regression Trees, CART: A User Manual for Identifying Indicators of Vulnerability to Famine and Chronic Food Insecurity*. International Food Research Institute, 1999.