Ostfalia
Hochschule für angewandte
Wissenschaften

Faculty of Supply Engineering

BACHELOR THESIS

Data Mining Techniques Applied to a Building-Integrated Hybrid Renewable Energy System

To obtain the academic degree

BACHELOR OF ENGINEERING

at the Ostfalia University of Applied Sciences, Wolfenbüttel

Bio- and Environmental Engineering

Author:	Catherine Fait
First supervisor:	Frank Klawonn
Second supervisor:	Corinna Klapproth/Ekkehard Boggasch
Submission date:	25 May 2018

Declaration

I hereby certify that this thesis is an original work authored by me under the guidance and advice of my colleagues and advisors. To the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where reference is made in the thesis itself.

By signing my name and the date below, I certify that what is written above is true.

Wolfenbüttel, 25. May 2018

Catherine Fait

Acknowledgement

I would like to thank the many friends, family members, colleagues, professors and administrative staff that have helped me get to this point. Your help, understanding and belief in me has been a major factor in my success and is greatly appreciated.

Contents

List of Figures

List of Symbols

A	sweap area of a turbine blade
c	scale parameter of the Weibull function
e	Euler's number
f	function of
K	efficiency coefficient of wind turbine
m	meters
P	available power [W]
ρ	rho, density of air
s	seconds
v	speed
W	Watts

List of Abbreviations

API	Application Programming Interface
COV	Change-Of-Value
CRISP-DM	Cross Industry Standard Process for Data Mining
HES	Hybrid Energy System
HREP	Hybrid Renewable Energy Park
HRES	Hybrid Renewable Energy System
LON	Local Operating Network
MySQL	My Structured Query Language
n. d.	no date
SQL	Structured Query Language

Abstract

There are many challenges associated with sustainable development, and one of the greatest resources available in modern times is data. On the other hand, one of the greatest challenges is producing sustainable energy that puts a stop to excessive energy-related carbon emissions. Uses of big data and data mining techniques have been showing promising results in the renewable sector and are gaining momentum. This paper explores the opportunities afforded and challenges encountered by incorporating a data-driven approach into renewable energy development at a small hybrid renewable energy facility. It takes a detailed look at wind energy generation data streams, and, through a case study, encapsulates the first steps of transforming the measurements into valued output and input of the power facility. The Cross Industry Standard Process for Data Mining is used as a framework to implement this study. Visualization results showed interesting patterns in the data and limitations to the planned analysis were uncovered during investigating the data. The data mining standard process proves to be an excellent framework in the scope of small power facilities for finding and documenting limitations, and building a foundation for innovation.

1 Introduction

Energy is a fundamental concept, both in the way the observable universe works, and in the way built society functions. With the onset of changes in the global climate due in part to carbon emissions, there is an incentive to reduce carbon emissions from energy production. One opportunity to do so is by switching from fossil fuels to renewable energy resources. Renewable energy resources have been on the rise in recent years as a sustainable alternative to fossil fuels [1]. In 2007, renewable energy comprised 12% of the global primary energy supply and is projected to be 35% by the year 2050 [1].

Along with more renewable energy infrastructure developments comes lots of data from monitoring and control sensors [3]. Furthermore, as computing power and cloud technologies have developed in recent years, there are shifts in the possible ways data can be collected, stored and analysed, and the value that can be created from data streams. The term data mining has come to describe the extraction of useful information from large amounts of raw data [2]. It is common in marketing and advertising, where applications of data mining are often used to gain better understanding of customer behaviours. Though, there are also important applications of data mining to gain a better understanding of medical or environmental solutions, among other applications. This work aims to explore the use of big data in the renewable energy sector.

As internet connectivity increases globally and information communication technology (ICT) advances, there is a stronger motivation and ability to integrate big data resources in meaningful ways into the renewable energy value chain. This is an active area of research and development in the energy sector, although there are some socio-economic, infrastructure, and technological constraints in this effort [3]. This paper explores the technological possibilities and constraints of big data in energy generation by applying the cross industry standard process for data mining (CRISP-DM) in a case study with data from a building-integrated hybrid renewable energy system.

1.1 The case for renewable energy

Throughout history, commercialized energy generation has fuelled major developments in civilization. Most recently, the advent of electrical power has driven artificial lighting, computer technology and the internet; most of modern society relies on connectivity to the electrical grid. Electricity generation currently relies on a portion of fossil fuels such as oil. Additionally, the immense transportation sector that is the backbone of global trade still relies heavily on fossil fuels.

These developments came with some critical caveats. Namely, they are changing environmental conditions that have sustained humanity thus far [7] by polluting the environment with toxic substances and releasing excess carbon into the atmosphere. Fossil fuels, such as coal, oil and natural gas, are energy sources that contain carbon from dead animal and plant material which has sedimented below the surface of the earth over many years [12]. Human burning of these fuels for energy releases ancient carbon into the atmosphere, which can be observed by comparing historical carbon isotopes ratios [12]. The isotope ratios in the atmosphere differ naturally from those in living beings, and, as atmospheric carbon levels rise, the ratio in the atmosphere has been observed to be shifting toward the ratio in organic tissue [12]. This is an indication that the combustion of fossil fuels is a cause of higher atmospheric carbon levels [12]. Combusted carbon takes the form of carbon dioxide (CO_2) in the atmosphere, which is a greenhouse gas that contributes to different climate change feedback loops [12]. Increasing CO_2 levels have been observed to have a strong correlation with global temperature increases [12]. Overall temperature increases can have different effects on local weather patterns in different places [12], and, while predicting exactly how these changes will occur is not an exact science, there is a large consensus that efforts should be made to reduce carbon emissions from human activities.

With a growing population and increasing demands from energy infrastructure, alternative energy production strategies and frameworks are needed to move away from fossil fuel combustion. Figure 1-1 ([4], p. 1) shows that energy consumption in the world is on the rise. Renewable energy sources such as wind, solar and hydro offer the possibility to meet increasing energy demands without adding more CO_2 directly to the atmosphere. Many countries have adopted economic policies to promote the production of renewable energy. Furthermore, renewable energy sources provide low cost energy and mitigate uncertainty in the supply of fossil fuels, because they are constantly replenished. It can be seen in Figure 1-2 ([4], p. 2) that, while consumption of coal, natural gas, and other liquid fuels is higher than renewables, renewable consumption is growing faster than other types of fuel.

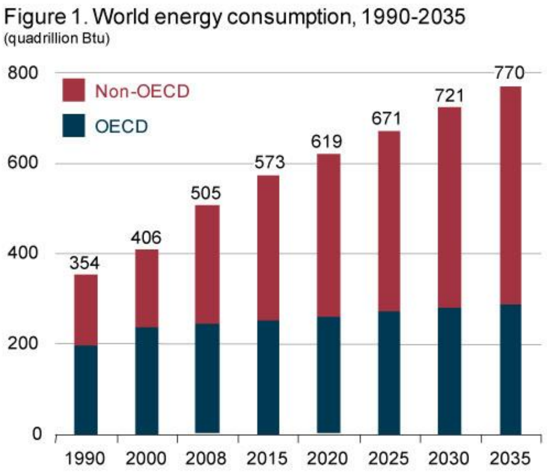


Figure 1-1: Energy consumption in the world from 1990 projected to 2035 ([4], p. 1)

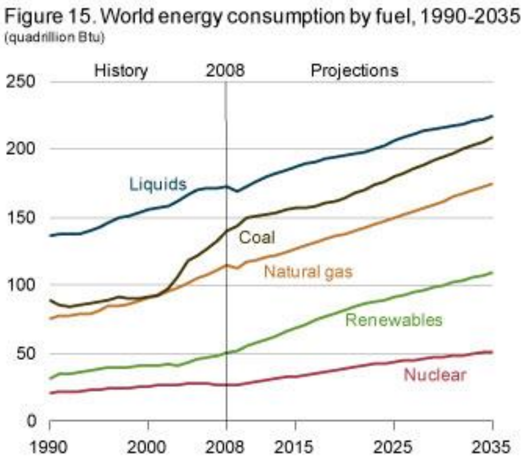


Figure 1-2: Energy consumption in the world by fuel type ([4], p. 2)

1.2 Aims of this work

There is a clear case for switching to renewable energy sources; however, there are many open questions associated with the implementation of renewable energy production. For example, instability due to unpredictable weather conditions, or complexity of combining several renewable sources make it hard to balance supply and demand and manage facilities. Due to recent advances in data technology, and the prevalence of sensors in renewable energy applications, large amounts of data can be collected and stored.

This work aims to take a preliminary look at what can be done with such data in a systematic way. It looks at energy data from a wind turbine in a building-integrated hybrid renewable energy system (HRES) test facility. It follows the cross-industry standard process for data mining (CRISP-DM) model as a framework to systematically discover how information insights and value can be generated from the facility's wind energy data. This will be done as an example case study, in which the operation of a decommissioned wind turbine is compared to that of its replacement.

2 Background on the renewable sector and data mining

The two key areas of background knowledge relevant to this work are renewable energy, specifically wind energy, and data mining. This section covers the relevant concepts in both these subject areas needed for understanding in depth the motivation and technicalities of this work. The first part of this section covers concepts of renewable energy and wind energy, the second subsection explains what is involved in the standard practices of data mining and how they are used in this work, and the final subsection explores the growing relationship between data and renewable energy.

2.1 Renewable energy

Renewable energy generation is based on the energy that reaches earth from then sun, is constantly replenished and does not leave any waste products. [7]. Renewable energy technologies include solar, wind, ocean (wave and tidal), hydropower, geothermal, biomass and biofuels, and hydrogen derived from renewable resources [7].

Renewable resources which are based on the weather, such as solar, wind and wave, are intermittent; therefore, there are challenges associated with integrating these resources into applications without compromising the reliability of electricity supply [22]. The principle problem is that when supply and demand of power do not match, either there is excess energy that needs to be directed somewhere other than consumer appliances, or there is not enough energy to fulfil the needs of the consumers. Figure 2-1 shows and example of a typical domestic power demand curve (left) and a renewable energy supply curve based on photovoltaic and wind sources (right), and how this miss-matched supply and demand scenario can be alleviated with electrical storage [22]. Electrical storage can help to level out the power peaks (a), to shift generated power in order to meet demands in times of low supply (b), or to compensate for the fluctuations in either demand or supply (c) [22]. The coupling of electrical storage with renewable sources is discussed further in section 2.1.2.

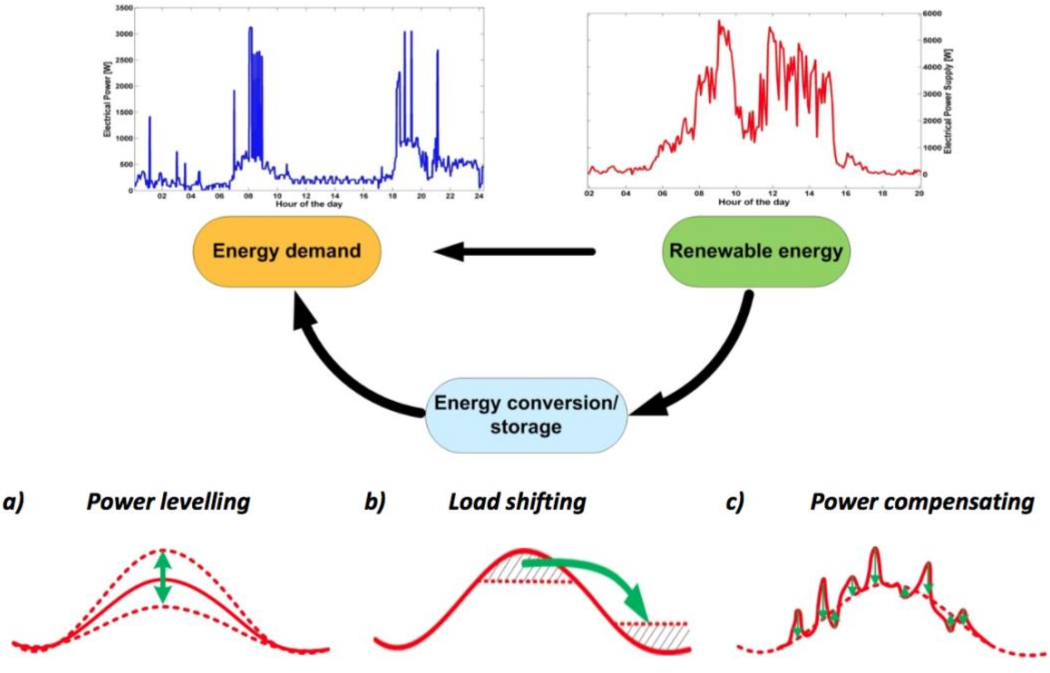


Figure 2-1: The miss-match of renewable power supply and power demand [22]

2.1.1 Wind energy generation and management

Wind energy is an indirect form of solar energy; essentially, currents are created in air due to thermal gradients from solar insolation [7]. As such, wind is always flowing, making it a readily available and completely renewable energy resource [7]. Wind, although following general patterns in the scope of the global climate [8] [9], is highly variable at any given point of interest, i.e. for energy capture. This means that the supply of wind is as unstable and unpredictable as the weather, which leads to unique challenges in meeting energy demand with wind power.

Wind energy is captured by turbines, which transform the mechanical energy of the wind spinning the turbine blades into electrical energy through a generator [7]. While turbine usage has been a recently growing trend for electricity generation, wind energy has been used for thousands of years, historically for pumping water and milling; the first uses of turbines date back to 500-900 AD in Persia [7]. Today, wind power is considered to have a high potential for electrical energy generation [10] and for increasing competitive sustainable energy production ([11], p. 10). Global electricity production from renewable sources is projected to rise from 18% in 2007 to 50% by the year 2050, and therein the portion of energy produced from wind shows a dramatic increase from the measured 2007 levels [1]. In 2013 in Germany, of the energy supplied from renewable sources (26%), the highest portion was supplied from wind (7.3%) [22].

Challenges with wind-based energy supply (or other weather-linked energy supply, such as photovoltaic) have to do with balancing the electricity supply and demand, with respect to grid capacity and the established operation of power production units [22]. In cases such as the German market, penetration of wind and solar energy sources have reached a point where production levels can be high enough that they need to be down-regulated in times of low demand [22]. This circumstance leads to wasted energy and is a technical challenge that needs to be solved [22]. Some practicalities and implications of the supply-demand management limitation of weather-based renewables is explored further in the next section.

2.1.2 Hybrid renewable energy systems (HRES)

A hybrid energy system (HES) is a platform for supplying energy that consists of different energy sources and storage, and the term hybrid renewable energy system (HRES) is used to refer to a HES based on renewable sources [22]. Definitions of HESs and HRESs are not precise in how they address scale and architecture [22]. Generally, the combination of different energy systems alleviates, to some degree, the limitations of individual components, and can increase reliability and efficiency [22] [13].

Due to the fact that renewables, such as wind, produce an inherently intermittent power supply, the application of energy storage technologies to temporarily decouple the supply from the demand can help to alleviate this problem [22]. Hybrid systems comprised of a variety of renewable sources with the addition of batteries, for example, would be more reliable and robust than a system relying only on a wind turbine. Wind and solar are commonly used together in hybrid systems.

One aspect in managing the increasing share of renewable energies is a shift from centralized, large-scale production to decentralized local or even building-level energy production [22]. This is especially relevant in isolated and developing places where energy infrastructure and connectivity to the grid is limited [13]. Hybrid systems comprising of renewable sources like wind and solar also have the added benefit of being inexpensive, and the current trend is that prices are decreasing [13].

Buildings are an important part of the energy ecosystem and there is interest in integrating HRESs into buildings with goal of more decentralized energy production [22]. Distributed hybrid systems provide the opportunity for low cost, clean, local energy generation and, can decrease the losses associated with transmission and distribution of electricity [13]. A test facility for a building-integrated HRES from which the data used in this paper have been acquired is described in the next section.

2.1.3 The hybrid renewable energy park (HREP) at Ostfalia University of Applied Sciences

The hybrid renewable energy park (HREP), located at Ostfalia University of Applied Sciences' main campus in Wolfenbüttel, is a test facility for researching building-integrated HRESs. It is a residential-scale distributed energy system consisting of photovoltaic panels, a wind turbine, combined heat and power, a proton exchange membrane fuel cell, an alkaline electrolyser, lead acid batteries and vanadium-redox-flow-batteries [22]. It also includes a programmable load simulator. This system was designed to represent a typical installation of such a system in a real residential application [22]. Figure 2-2 [22] below is a simple schematic showing different nodes of the HREP at Ostfalia. This work is mostly concerned with the data generated at the wind turbine node of the HREP.

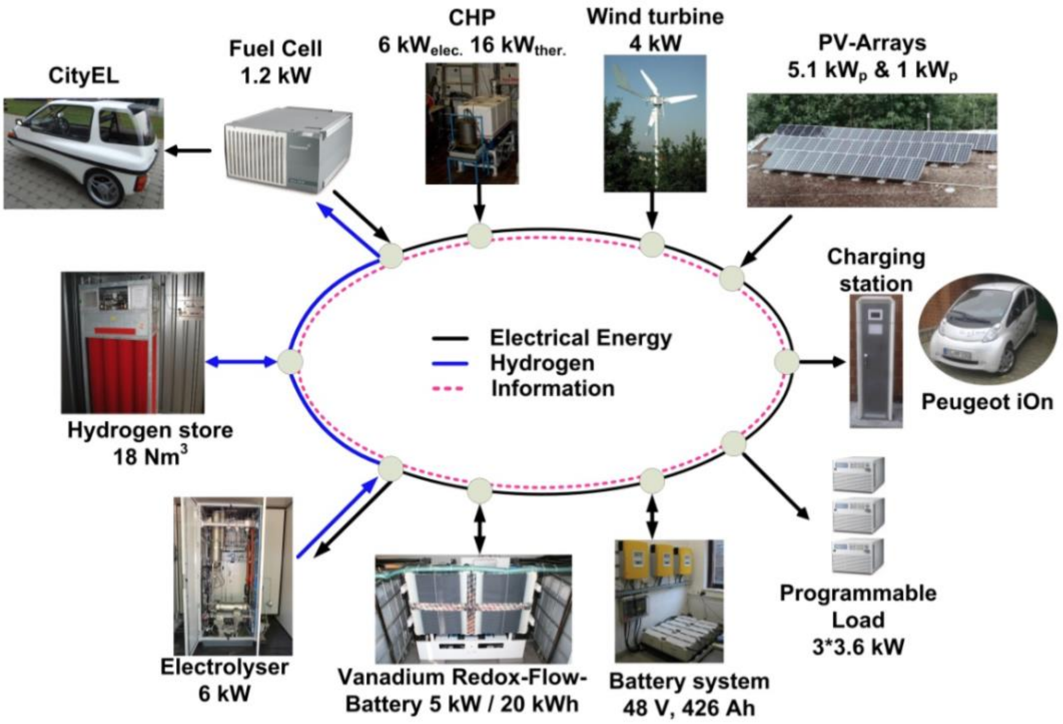


Figure 2-2: Simple schematic of the HREP at Ostfalia [22]

2.2 An overview of data mining

Data mining has come to have several definitions. Generally, data mining is the acquisition of useful, actionable information from large datasets with the help of computing power through various intelligent analyses [17] [2]. Techniques from statistics, computer science, and importantly, domain knowledge from the field that is subject of the analysis all come together to make this happen. In principle, data mining can be applied to any project which produces large volumes of data.

This section covers briefly the development of the field of data science and the concept data mining. The aim is to deliver an understanding of how the usage and value of data is evolving, as to provide a better assessment of the role of data in the renewable energy industry. Furthermore, a standard practice for data mining is defined and its role in this work is discussed.

2.2.1 Data in the information era

Over the last half a century, there have been several important developments that have changed the way information is collected, processed and shared. This section takes a brief look at those developments, namely the invention of the computer and the internet. The computer allows for the storage of information digitally, while the internet allows for that information to be sent across the globe instantaneously. This means that more data is becoming increasingly more accessible.

Digitalization, the switch form paper records to digital records is still today a topic of interest for many organizations. Not only has record keeping seen a change from digitalization, but also have the processes of taking measurements themselves, i.e. with digital meters [14]. Moreover, internet usage rates have increased 1052% globally between 2000 and 2018, and on average globally, one in two persons are internet users [5]. These developments make it possible to know a large amount of information about almost anything almost instantly. For example, satellites take vast amounts of measurements of the earth and this data is used in many applications, from military to meteorology to open datasets such as the CORINE Land Cover inventory [15], and Google's Map application programming interface (API) [16]. More and more companies are making their data publicly available to individual developers through APIs.

In the current situation, relevant questions are, what to do with this onset of large amounts of data, and what affordances do current technologies allow? The next section examines the concept of data mining and how it's distinction from other related fields can help provide an answer to this question.

2.2.2 Distinction from other fields

Data mining is a concept from the field of data science. Although not a new field per se, data science has become more popular in recent years, particularly as many industries gain access to and benefit from very large data sets (big data). Data mining and statistical analysis have many overlapping concepts, however there is a fundamental difference. Statistical theory was developed prior to computers, and therefore strives to make inferences about a population based on small samples, while data mining employs computational power and digital memory and affords to analyse data sets that can contain entire populations and a vast number of variables [2]. As such, some differences in approach from traditional statistics are required [2]. For example, statistical techniques may lead to overfitted models or results in which many variables have statistical significance, or otherwise unhelpful results [2]. Moreover, in data mining projects, data are not necessarily collected for the purpose of answering the given question [18], while in statistics, studies are designed and data collected with the end goal in mind. Further, data mining aims to not only show significant results, but also evaluates the results in terms of generated value to the project or business [2]. Therefore, an understanding of the domain is important to decide which variables to consider, in which way, and why.

There are different tasks that data mining can accomplish, and projects are unique. The data mining tasks are overviewed in the next section.

2.2.3 Data mining tasks

There are different ways the product of a data mining project can look in general, and each individual project takes on its own unique form solution. Some common data mining tasks are *description*, *estimation*, *prediction*, *classification*, *clustering*, and *association* [17]. These six tasks are briefly described in this section.

Description is a task which aims to describe patterns and trends in the data. It also aims to visualize these patterns and trends in a concise, coherent, and transparent way [17].

Estimation, also known as *regression*, takes a number of variables, and based on their values, predicts an unknown *numerical* variable [17].

Classification is similar to estimation, although the target variable is categorical rather than numerical [17]. Classification is a popular data mining task.

Prediction is similar to estimation and classification, although the unknown variable lies in a different time (i.e. the future) than the variables used to predict it [17]. Estimation and prediction are also popular.

Clustering, also known as *segmentation*, is the grouping of data objects with other data objects that are similar in value or composition [17]. It may be that some data objects are not similar to any others, and these are considered outliers. Classification can be used to detect anomalies and outliers and to define the general structure of the data, and may be also useful to partition the data set in smaller subsets for further analysis [17].

Association is an analysis which quantifies relationships between attributes of data objects, such as defining a frequency that data objects with a given attribute value also have other given attribute values [17].

The implementation of these tasks is discussed in the next section.

2.2.4 The CRISP-DM model and its role in this paper

The Cross-Industry Standard Process for Data Mining (CRISP-DM) defines a general model of a data mining project. It was developed in 1996 [17] [2] and has been widely adopted since then. It comprises of 6 phases throughout the lifecycle of a project, namely *business understanding*, *data understanding*, *data preparation*, *modelling*, *evaluation*, and *deployment*. Rather than following a linear progression through the timeline of the project, the phases are iterated through adaptively, based on accomplishments of the previous phase. This relationship is illustrated in Figure 2-3 (adapted from [18], p. 9). The six phases are described next.

Business understanding is also called *project* understanding or *research* understanding. This is the first phase and initializes a project. In this phase, the objectives and requirements and resources of a project are identified; these are then prosed into a data mining problem definition, and a preliminary plan for solving it is created [17]. Instead of the other three terms for this phase, the term ***value understanding*** is used in this work, as it covers all three of these terms and underscores the fact that creating some value for the data owner or society – monetary or non-monetary – is the goal of a project.

Data understanding is the second phase, and it is when first contact with the data happens. This includes collecting the data, evaluating its quality, and using exploratory data analysis to find expected and unexpected patterns in the data [17]. During this phase, an assessment is made concerning the suitability of the data available to solve the problem ([18], p. 9). If it is not suitable at all, the project may be cancelled, or, if it is partially suitable, the objective should be revised, show in Figure 2-3 (adapted from [18], p. 9).

Data preparation, the third phase, consists of fixing data quality issues and transforming the raw data into an appropriate composition for modelling [17]. For some analyses, new attributes need to be generated ([18], p. 9).

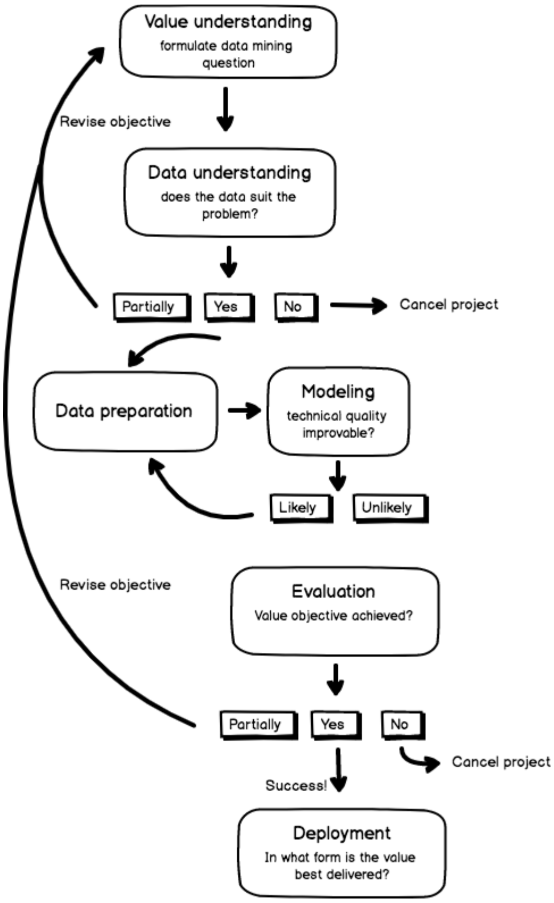


Figure 2-3: The CRoss Industry Standard Process for Data Mining (adapted from [18], p. 9)

Modelling is the part of the process where the data are utilized to gain new information. In this phase, it is important to select the appropriate modelling techniques for the problem, and there may be more than one [17]. Jumping back and forth between the modelling and data preparation phases is not unlikely [17], as different modelling techniques have different needs, and the technical quality of the modelling may be improved via better data preparation ([18], p. 9).

Evaluation is as important as modelling. In this phase, the models are tested to see if they *do* generate new information and answer the data mining problem identified in the first phase ([18], p. 9). The model is run as a test to see if it will in fact answer the initial data mining problem statement, before deploying it in the field. Different data mining tasks have different evaluation methods and formulas [17].

Deployment of a model on the data is when the real insights with value are generated. This phase is essentially the implementation of the model(s) that passed the evaluation, and the deployment may take many different forms in different industries and projects [17]. For example, it could be continuously integrated to analyse newly collected data or a report of the project findings.

It should be noted that the six phases are ordered chronologically, and adaptive iteration through the phases only happed backwards. For example, one might skip back from the evaluation phase (5th step) to the business understanding phase (1st step); however, it would not make sense to skip forward from the business understanding phase directly to the evaluation phase, because there would be no model yet to evaluate. Therefore, this process model can be described as highly reflective, where an analyst often reflects on their work up to the current point in order to decide on the next steps.

The solution presented in this paper is organized chronologically based on the six phases of CRISP-DM. The aim of this paper is to follow the above outlined process to generate new insights for the HREP from its sensor data. At a high level, the objective of this work is exploration of value propositions of data mining techniques at the HREP. For that purpose, following the CRISP-DM structure provides a strong foundation in industry standard practices, giving the more abstract types of results in this work integrity. At a practical level, this work is concerned with a case analysis and employs the CRISP-DM structure as a working plan throughout the case.

As the phases of the CRISP-DM process are adaptive and may be revisited, the end of some chapters include a summary of any revisions of previous sections or of the project objective.

2.3 Big data and data mining in the renewable energy sector

Information and communication technology advancements mean that digitalization is relevant throughout the entire energy value chain, and large amounts of data from energy asset monitoring and control systems are becoming available [3]. There are opportunities to improve the operation of energy systems within the context of both demand and supply by utilizing techniques used to mine big data [3]. For example, in the scope of demand, data mining can provide an insight into consumer behaviour to build better load profiles; in the scope of supply, data mining techniques could be used to build better fault diagnostics and scheduling of maintenance operations [3].

There are two different scenarios to consider when it comes to the applications of big data in the renewable energy sector. On one hand, there are large centralized electrical power systems which facilitate complex supply and demand networks via smart control centres [6]. On the other hand, there are small distributed hybrid systems which have many different technologies and dependencies, and therefore unique optimization needs [22]. The latter is mainly discussed in the next section. Nevertheless, one central theme in renewable energy across all systems is the usefulness of big data for optimizing tasks related to power generation, transmission, distribution, and demand side management [3].

A case example of using big data techniques in renewable energy is outlined in [3]. In this study, a data-driven approach is applied to wave power site assessment and the integration of wave energy with maritime developments [3]. The study makes use of the BigDataOcean project's innovative value chain of maritime data streams [19] to empower the development of wave energy.

3 Building a unique value understanding of the HREP

This and subsequent sections cover the data mining project itself, in the form of a case study using the CRISP-DM process to compare a decommissioned wind turbine with a new one. Also discussed is the availability and the value of data resources at the HREP facility. As this is a first look at applying the CRISP-DM process in the facility, there is an interest to assess how this process can facilitate future insights generated from the HREP data. Therefore, a focus is placed on understanding the quality and accessibility of the data. Then, possible deployment solutions are discussed in the last chapter.

The first phase in the CRISP-DM process is developing an understanding of the project's unique needs and resources. This section is dedicated to applying the previously discussed understanding of the energy industry, HRES operation, and the HREP at Ostfalia to formulate a technical data mining question and clarify the value created.

There is a large amount of data being generated from the sensors at the HREP. Figure 3-1 [22] is a diagram of the electrical layout of the park, and Figure 3-2 [22] shows the communication topology. Relevant to this analysis is that two wind speed sensors, a wind direction sensor, and a wind power meter, which are connected via a local operating network (LON) to an SQL database.

In the past, various computer simulation strategies have been implemented at the HREP to design, assess, and develop control strategies for the system [22]. Currently, there is an interest for assessing the operational differences between a recently installed turbine with the previously decommissioned turbine. Furthermore, due to the inherent imbalance of supply and demand of renewable energy discussed earlier, gaining a deeper understanding of the supply characteristics of the energy park is valued.

The data mining question(s) can be formulated as: is there a power output difference between the old and new turbines during months with similar average wind speeds? Furthermore, it is of interest to know, how is wind power correlated to wind speed and wind direction for both turbines? Lastly, how can further insights be generated through the collected data at the HREP in the future?

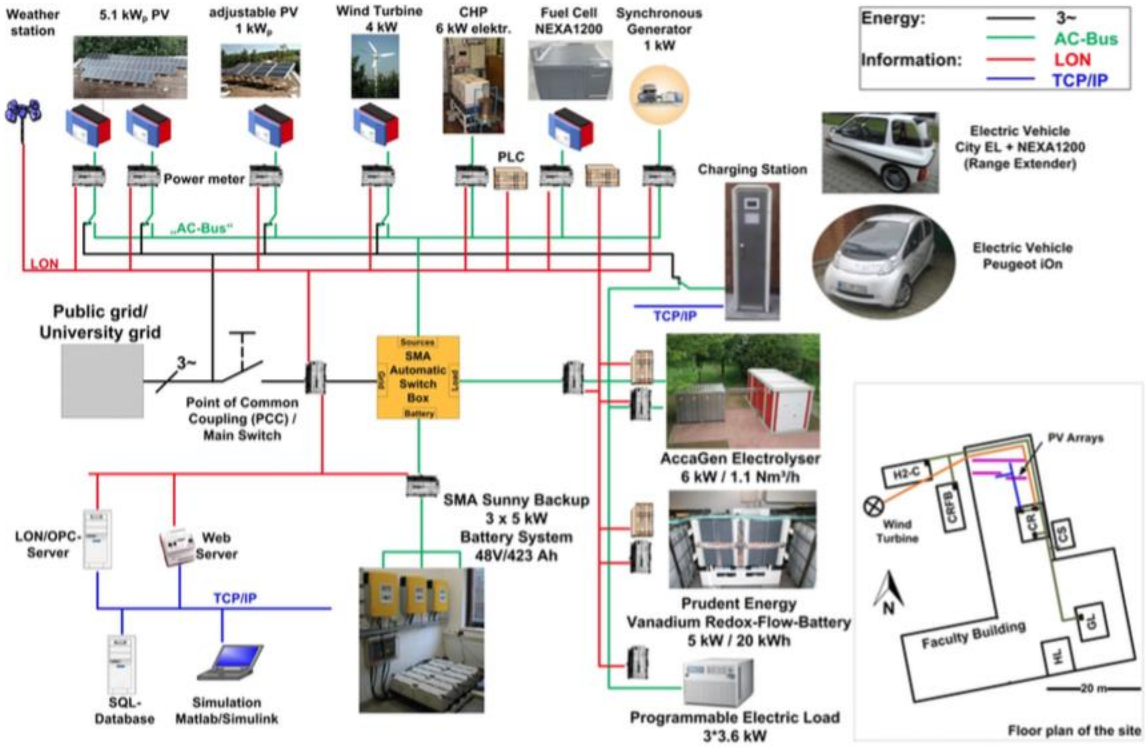

Weather station
5.1 kWp PV
adjustable PV 1 kWp
Wind Turbine 4 kW
CHP 6 kW elektr.
Fuel Cell NEXA1200
Synchronous Generator 1 kW
Energy: 3~ AC-Bus
Information: LON TCP/IP
Electric Vehicle City EL + NEXA1200 (Range Extender)
PLC
Power meter
Charging Station
„AC-Bus“
LON
Electric Vehicle Peugeot iOn
Public grid/ University grid
3~
TCP/IP
Sources SMA Automatic Switch Box Battery PV Load
Point of Common Coupling (PCC) / Main Switch
AccaGen Electrolyser 6 kW / 1.1 Nm³/h
PV Arrays
H2-C
CRFB
CR
CS
Wind Turbine
LON/OPC-Server
Web Server
SMA Sunny Backup 3 x 5 kW Battery System 48V/423 Ah
TCP/IP
Prudent Energy Vanadium Redox-Flow-Battery 5 kW / 20 kWh
N
Faculty Building
GL
HL
20 m
SQL-Database
Simulation Matlab/Simulink
Programmable Electric Load 3*3.6 kW
Floor plan of the site

Figure 3-2: Simple electrical schematic of the HREP [22]

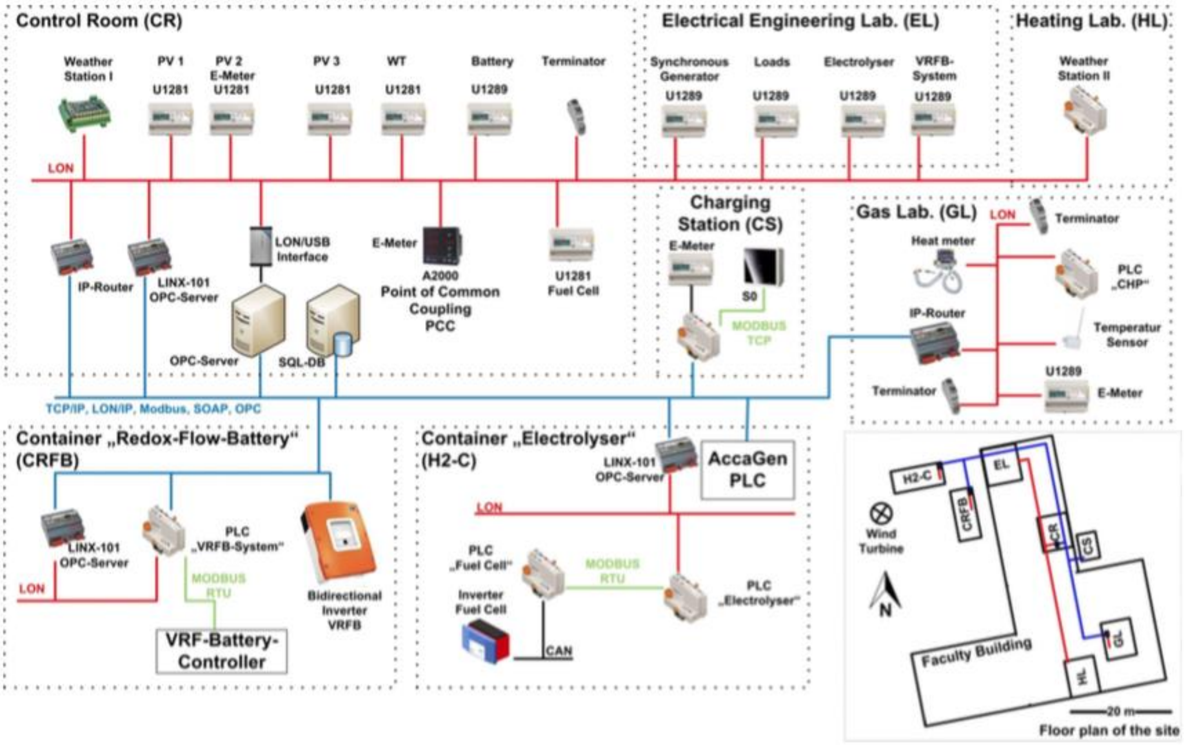


Figure 3-1: Communication and control schematic of the HREP [22]

3.1 Data needs at the HREP

This section builds a deeper understanding about the domain value of the data from the HREP, through earlier works.

In a system with many interdependent nodes like the HREP, a complex network of sensor and meter data streams exists, and this network needs to facilitate the management of energy flows within the park [22]. Analysing power flow accurately is an important aspect of this management, as power variations need to be compensated by the hydrogen system (Figure 2-2) [22]. An important issue is the temporal resolution (granularity) of the data. A high temporal resolution of power data is desired, of minutes or even seconds, because averaging over long time intervals can lead to underestimating power peaks [22]. Figure 3-3 [22] shows the effect of averaging photovoltaic power data over different temporal resolutions, where averages over one hour drastically underestimate the power peaks.

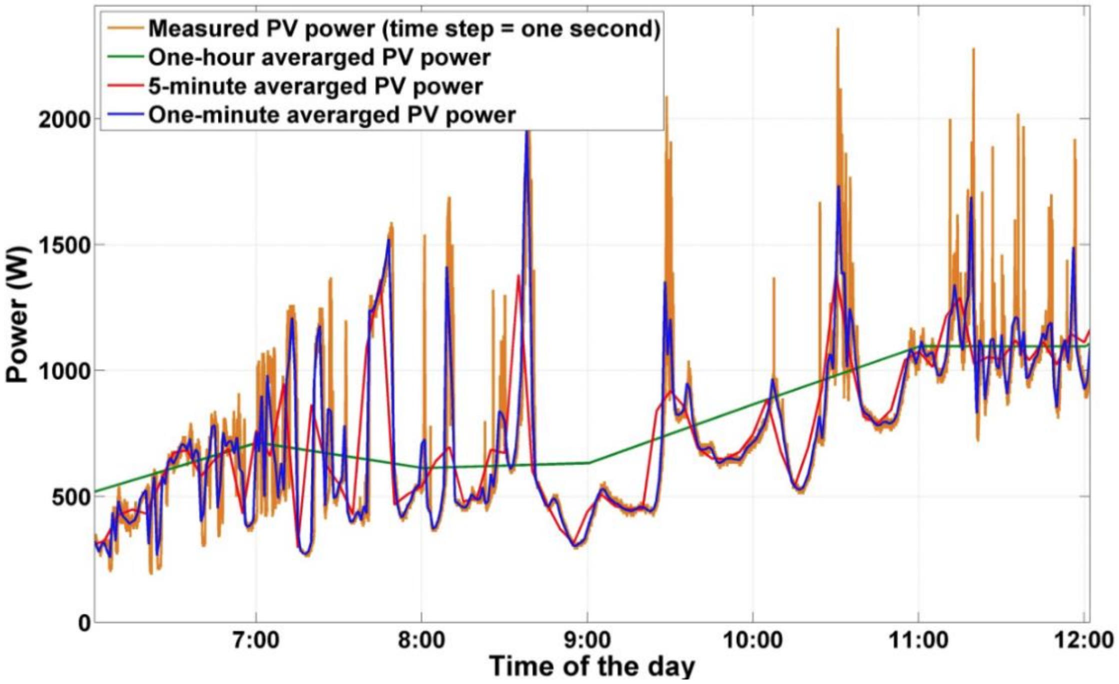


Figure 3-3: Photovoltaic power data from the HREP at different temporal resolutions [22]

3.2 Comparing the efficiency of the old and new turbine

Describing the behaviour of the old and new wind turbines at the HREP is a primary aim of this analysis, and this section discusses the relevant details of the events concerning the wind turbines. This is a descriptive data mining task, in which a comparison of monthly average wind speeds is first conducted, and then a comparison of monthly average power and energy outputs is compared in months with a similar average wind speed. Furthermore, correlations between the wind power, wind speed and wind direction are also of interest.

In 2017 a new wind turbine was installed at the HREP. The old turbine (Geiger SG 500/ ASP inverter, 4 kW / 3.6 kW, 2003 [22]) went out of service in November 2016. The new turbine was installed in August 2017 and started to work in October 2017. Therefore, from December 2016 – September 2017 inclusively, there is presumably no data, and for the months November 2016 and October 2017, there may be a portion of the records.

3.3 Assessing the available data resources

As this work is a preliminary look at using data mining techniques and the CRISP-DM model within the analysis at the HREP, an important consideration is how it can set the foundation for future work with such techniques at the HREP. This adds value to the project beyond the scope of a single analysis. Therefore, in the next chapter (data understanding) the accessibility and quality of the data are described, and the implications of this are discussed in the deployment section. Where appropriate, frameworks or methods are defined to retrieve the data in a specific format, and recommendations are made for future work.

4 Data understanding

This chapter covers the important data understanding phase, which includes acquiring the raw data, pre-processing it, as well as preliminary exploratory data visualizations to better understand the structure and quality of the data. Pre-processing includes cleaning the data of, for example, any missing or incorrectly entered attribute values, identifying missing records or outliers, and any other data quality issues that could impact the ability to perform the analysis [17]. Visualizations include various plots to check for general trends or anomalies. At this point, a thorough understanding of the domain obtained in the value understanding phase is an essential building block for the project.

The idea of data quality refers to how well the data match with their intended use [18]. It is important to note that data contain bias, and it is often the case that the data were not collected for the purpose of answering a specific data mining question [18]. For example, data with a high population of a particular subgroup might deliver a result which is biased toward that subgroup. In a statistical study, the sample would be collected in a way to avoid this effect. However, in a data mining project where the data is already collected, data need to be handled as they are, as designing unbiased data collection can only be done for the future, and is unrealistic in some cases [18]. Data quality is important to consider early in the project so that during the data processing phase the effects of biases and inaccuracies can be addressed.

In the case of the wind data, they are collected by sensors and meters over time. The sensors, when functioning correctly, have a defined amount of measurement inaccuracy. The accuracy of the power meter for the wind turbines is 1% ± 1 digit; the accuracy of the wind speed measurements is $\pm\, 0.3\ \frac{m}{s}$; and the accuracy of the wind direction data is ± 2.5° [22]. Accuracy statements of sensors and meters not related to wind measurements are available in [22]. Other inaccuracies could come from malfunctions in the sensors, for example, and plotting the data can reveal such quality issues.

4.1 Analysis methods

The data are assessed using the R programming language, an open source statistical programming language. Code written for these analyses is attached in Appendix A 1. Code files are separated by their respective task; i.e., functions that used to import the data are stored in a file called `1_import.R`, functions used to clean and format the data are stored in a file called `2_clean_form.R`, and functions used to generate plots are stored in the file called `3_plot.R`. A main script calling these functions is stored in a file called `4_do.R`. The numbers in the file names represent which order they are to be executed. Code is organized this way to enable reproducible results from other wind data by just replacing input files and variable names. Steps explained in the following subsections are briefly described inside the code files with commented lines.

4.2 Data sources and description

Wind speed, direction and power measurements are collected continuously at the HREP, along with other sensor and meter measurements relevant to the different nodes in the system (Figures 3-1 and 3-2). These data are stored at the facility in an SQL database. The data are uploaded to the database every month, and therefore database files are separated by month. Data are logged to the SQL database on a Change-Of-Value (COV) mechanism, where a new value is recorded and time stamped only when there is a change of value [22].

The wind speed measurement unit is meters per second $[\frac{m}{s}]$, the wind power measurement unit is Watts $[W]$, and the wind direction measurement unit is degrees (0-360) [°]. Accuracy of the measuring devices are noted in section 4.

The wind power and speed data were extracted and pre-processed with a MATLAB tool created at Ostfalia which interpolates the SQL records into one-second time stamps. The wind direction sensor and a second wind speed sensor were installed after the creation of this tool, and these data are therefore only available in the COV format directly from the SQL database.

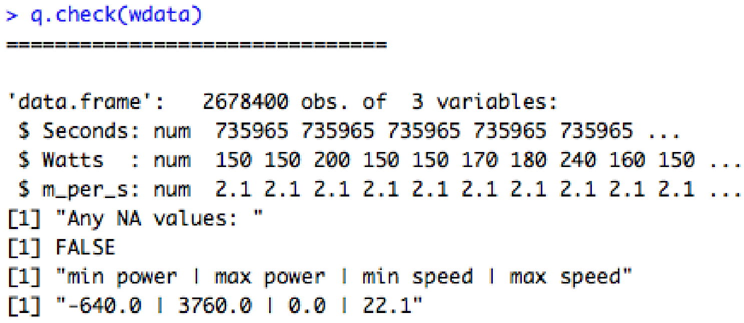
A local test MySQL server instance was run to open a connection to the database files, demonstrating direct accessibility via R. However, to compare the data already extracted with the existing MATLAB tool to the data from the wind direction and second wind speed sensors, there is a need to either modify the existing MATLAB tool or to recreate a new interpolation/extraction tool with R.

4.3 Data quality assessment

In this section, the quality and structure of the data are examined for a prototype month. The accuracy of data is a measure of how well the data represent the truth [18]. Domain knowledge acquired in the previous phase can help discern what the true values *might* be. Since the data have already been pre-processed with the MATLAB tool prior to acquisition, it is more likely that the records are complete and syntactically accurate, though this does not mean the values are representative and an exploratory quality assessment is still needed. As the data are extracted from the database using a MATLAB tool, the corresponding `` `.mat` `` data file was transformed into a format readable by R using the R package `` `R.matlab` `` [20].

First, the relevant columns of power, speed and seconds were taken as a working subset of the data. Records were checked for completeness, and syntactic and semantic accuracy. Regarding completeness, since the data are interpolated to values per second, the month of January should have 2678400 seconds (equation 4-1), as it does, shown in Figure 4-1.

$$60\frac{seconds}{minute} \times 60\frac{minutes}{hour} \times 24\frac{hours}{day} \times 31\frac{days}{month} = 2678400\ \frac{seconds}{month} \tag{4-1}$$



```
> q.check(wdata)
================================

'data.frame':   2678400 obs. of  3 variables:
 $ Seconds: num  735965 735965 735965 735965 735965 ...
 $ Watts  : num  150 150 200 150 150 170 180 240 160 150 ...
 $ m_per_s: num  2.1 2.1 2.1 2.1 2.1 2.1 2.1 2.1 2.1 2.1 ...
[1] "Any NA values: "
[1] FALSE
[1] "min power | max power | min speed | max speed"
[1] "-640.0 | 3760.0 | 0.0 | 22.1"
```

Figure 4-1: Preliminary data quality check results (Jan 2015)

The data is also checked for "NA" attributes, non-numeric attributes, and maximum and minimum values. Figure 4-1 shows the results of this check. The prototype month in this example is January 2015. A wind speed of less than 0 is not expected, however it may be that power can take negative values. This is because when wind speeds are too low, the turbine inverter is in stand-by mode and a small amount of power is consumed (personal communication, E. Boggasch, 23 May 2018). The cut in value for the inverter to turn on in about 80 Watts (personal communication, E. Boggasch, 23 May 2018).

This is a basic quality check of the data, and further visual explorations are covered in the next section.

4.4 Exploratory visualizations

Visualization is a powerful tool to help understand large amounts of information within minimal amounts of time and space. In this section, the data are plotted in various ways with various granularities to identify known and unknown patterns, outliers, deviation trends, and any other notable issues that can be comprehended visually.

A simple preliminary plot to check for missing values is a time series plot. Figure 4-2 and 4-3 show how wind speed and power, respectively, change over time in January 2015.

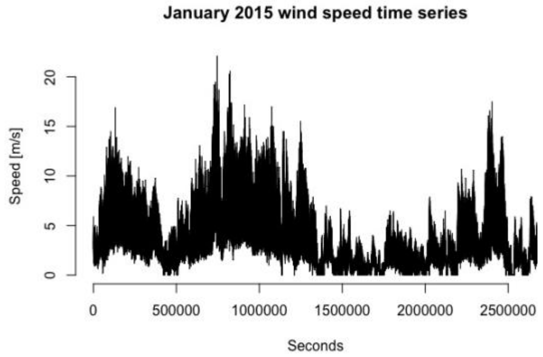


Figure 4-3: Wind speed time series for January 2015

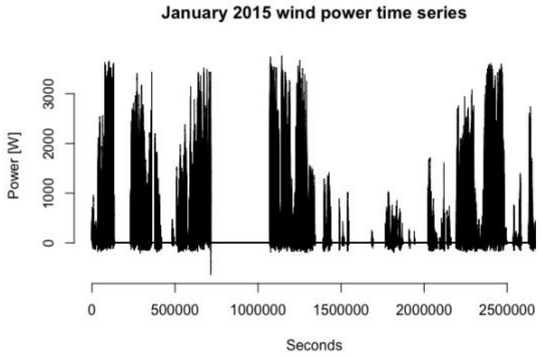


Figure 4-2: Wind power time series for January 2015

However, looking at January of the following year, 2016, shown in Figure 4-4, there is a noticeable deviation in the data, where the wind speed plateaus at around 3 $[\frac{m}{s}]$ for about a week. This can be attributed to a period when the turbine was out of service, and the interpolation tool has caused the last recorded value to persist in the records until data was being generated again. This kind of deviation in the data will affect the planned comparison of the monthly averages.

Another thing to check for is the distribution of values in the data. Wind speed follows a probability distribution called the Weibull distribution, given by the following formula (4-2),

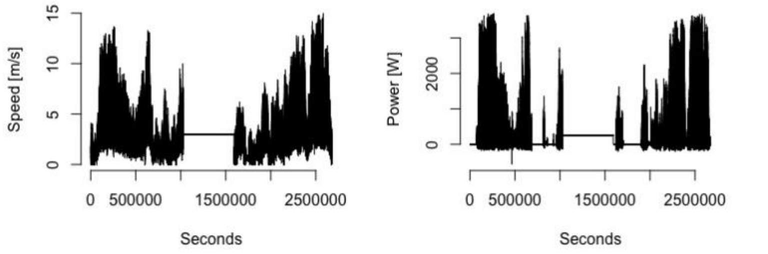
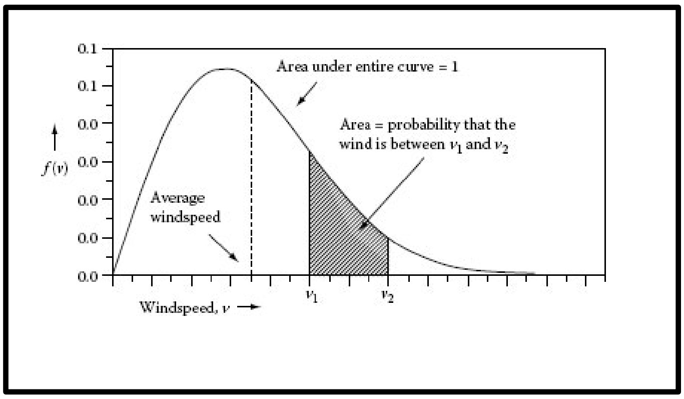


Figure 4-4: Wind speed and power time series for January 2016

$$f(v) = \frac{2v}{c^2}\, e^{-(\frac{v}{c})^2} \qquad (4\text{-}2)$$

where v is the wind speed, and c is a scale parameter that varies by region [7].



0.1
0.1
0.0
0.0
0.0
0.0
0.0
f(v)
Area under entire curve = 1
Area = probability that the wind is between v_1 and v_2
Average windspeed
Windspeed, v →
v_1
v_2

Figure 4-5: The Weibull probability density distribution [7]

Figure 4-5 [7] shows a plot of this distribution. Comparing this expected behaviour to the frequency of measured wind speed values in Figure 4-6, it can be observed that the data generally follow this pattern, except for one bin around 3 $[\frac{m}{s}]$ in January 2016. This is the same value for which the time series plots show missing data, and these data are therefore not an accurate representation of the true wind speed. The power time series plot shows the same behaviour in the time period in question, so it can be concluded that either both measuring devices were malfunctioning at the same time (unlikely) or the turbine was not running or some other systematic event occurred.

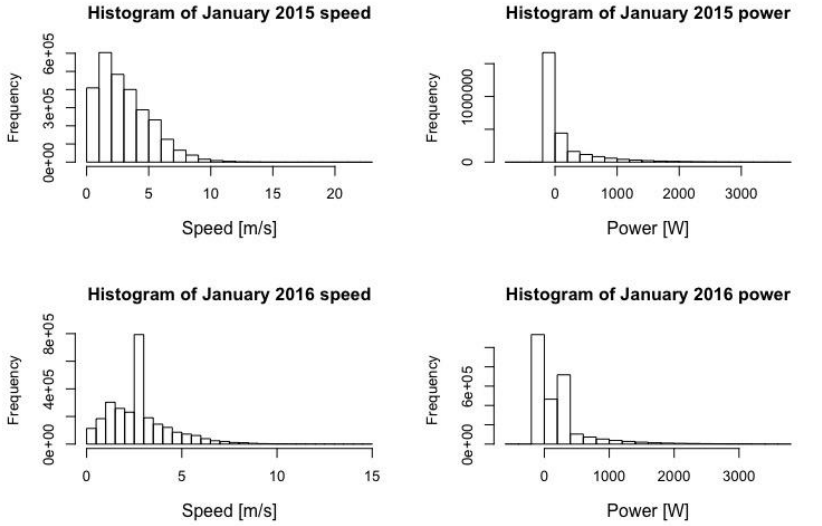


Figure 4-6: Histograms of speed and power values in January 2015 and 2016

It can also be observed that the highest frequency of values for wind power appear to be below zero. This seems unusual; however, compared to the time series plots of wind power, there are indeed lots of near zero measurements. Knowing that the power inverter is in standby mode at low wind speeds, and, observing from the wind speed histograms that the highest frequency of wind speed measurements is between 0 and 5 $[\frac{m}{s}]$, these kind of power values seem to make sense and can be accepted for further analysis. One thing to note is that the tails of the histograms stretch out quite far, and the higher values in the observed range could have very low occurrences and be deemed as outliers. One way to better visualize this is with a box plot, shown in Figure 4-7, where the box shows the middle 50% of the data (the inter-quantile range), the whiskers are 1.5 times the inter-quantile range, and any data points beyond that are drawn as points [18]. The figure shows that the wind data are gathered more closely together, with some outliers in the higher values, while the power data is distributed more about the observed range.

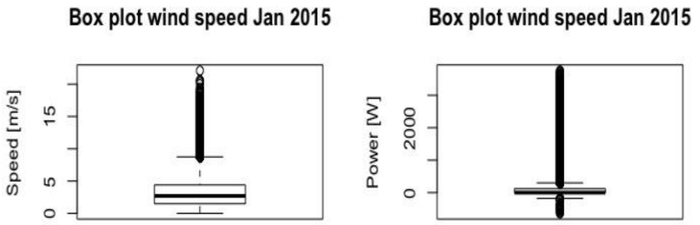


Figure 4-7: Box plots of wind power and speed for January 2015

As mentioned in [22], it is also interesting to examine the data at different granularities. Averaging over a longer period of time will reduce the noise in the data, however it also masks power peaks that are important for management of the facility. While hourly averaging is not recommended in [22], here it is explored to look for any monthly trends. Figure 4-8 is a time series analysis of both the power and speed in January 2009. What can be seen from this plot is that the power curve follows the trend of the speed curve. This granularity allows to look, for example, for daily or weekly trends, but may mask missing values.

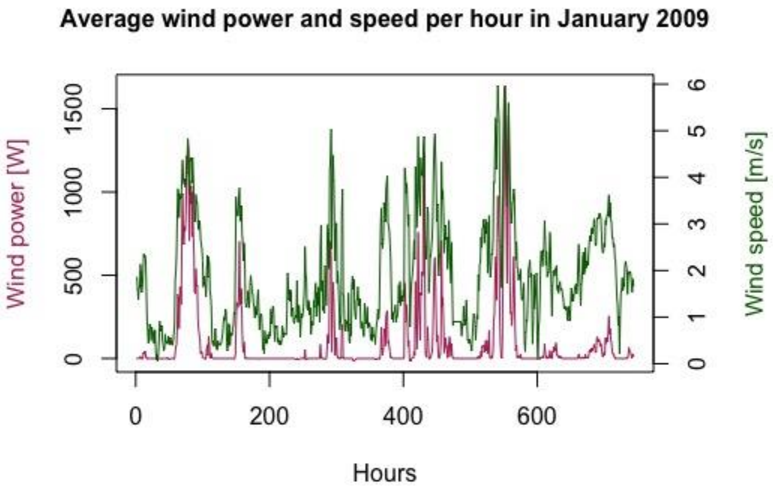


Figure 4-8: Hourly averages of speed and power in January 2009

Another interesting relationship to look at in this phase is the correlation between wind power and speed. This relationship is also examined at different granularities in Figures 4-9, 4-10, and 4-11.

Figure 4-9 shows the correlation between wind power and speed on the first day of January 2009. Hourly data is taken every 6th hour. During some hours, a relationship can be distinguished visually, while some hours contain very few changes in values. The cross-correlation is shown beside each graph to examine the lag between the two series. The cross-correlation method used can have sensitivity to outliers (personal communication, F. Klawonn, 23 May 2018), and these visualizations are meant as preliminary exploration only. Nevertheless, from the box plot examined in the previous section, there is yet no evidence of extreme outliers. When the cross-correlation function shows a peak at 0, the series can be said to be auto-correlated. Peaks appear when the functions are in the same phase. Figure 4-9 does not make a strong case for identifying a lag, though in two of the plots, there appears to be a lag of 10 time units (seconds).

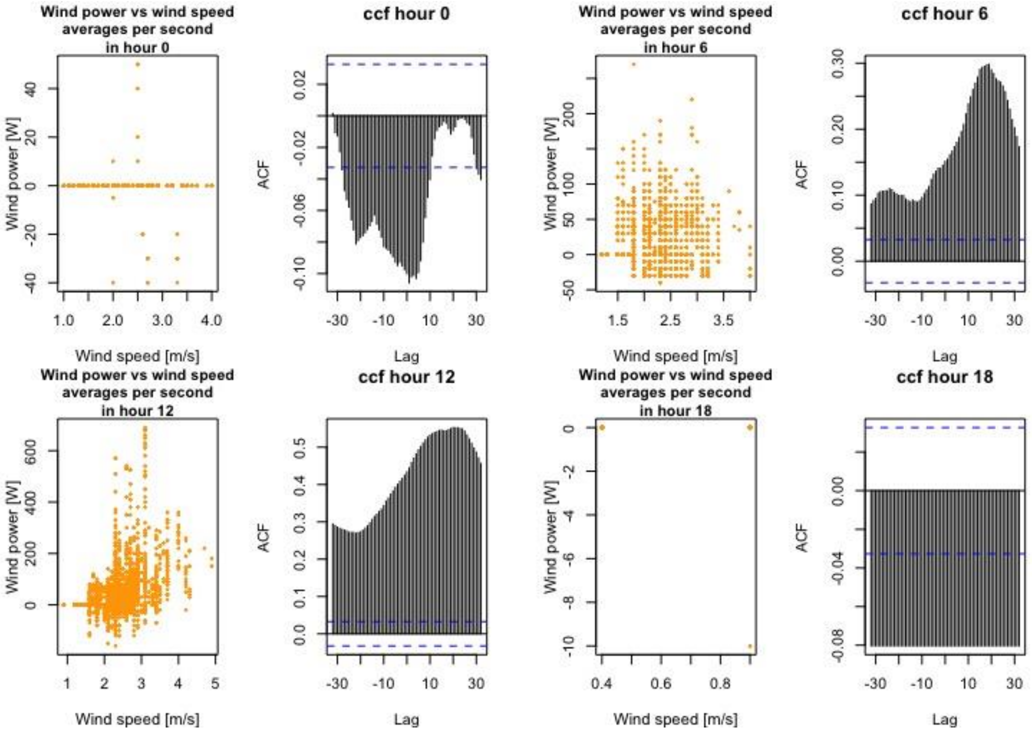


Figure 4-9: Correlation of wind power and speed on one day in January 2009

Figures 4-10 and 4-11 represent the granularity of minutes and hours, respectively. Figure 4-10 shows the minutes in a random sample of days in a month, while Figure 4-10 represents all the hours in the month. As the data have been smoothed over the averaging process, the series appear to be more auto-correlated and less lag is observed. Also, the lower the granularity, the more the plots resemble the cubic theoretical relationship between wind speed and wind power, given by the equation (4-3),

$$P = \frac{1}{2} A\rho v^3 K \qquad (4\text{-}2)$$

where P is the available power, A is the sweap area of the turbine blades, ρ is the density of air, v is the wind speed, and K is the efficiency coefficient of the turbine [21]. The black points in figure 4-11 depict a cubic function with respect to speed.

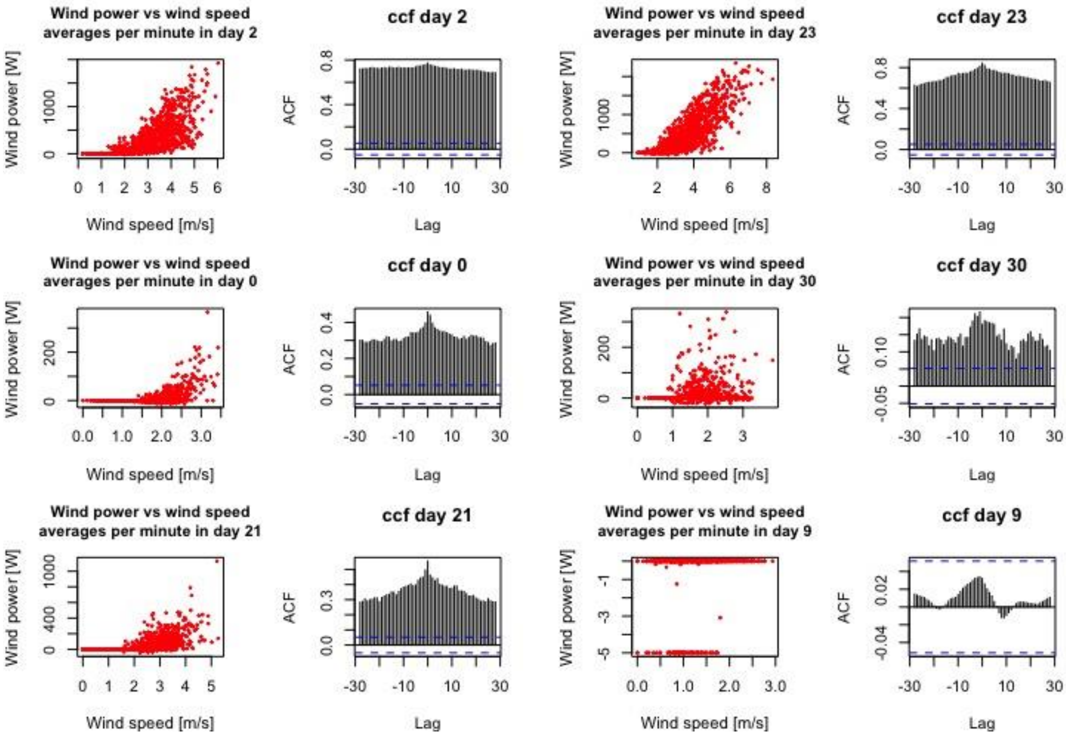


Figure 4-11: Correlation of wind power and wind speed in minutes of random days in January 2009

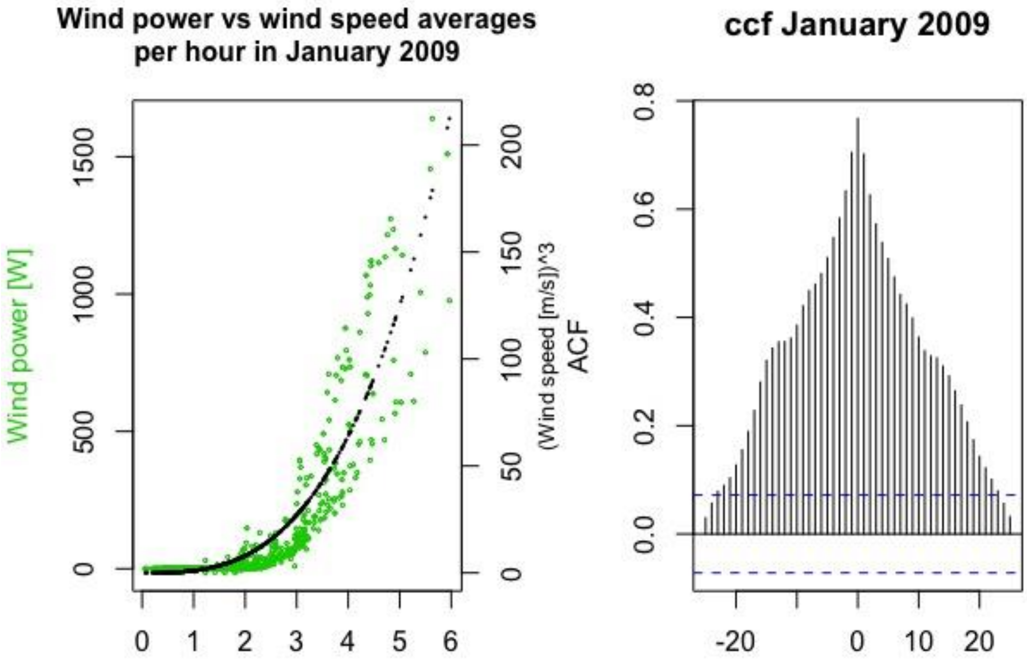


Figure 4-10: Correlation of wind power and wind speed in hours of January 2009

4.5 Data understanding phase summary

The accessibility of the wind direction data is low, despite the proof of concept for accessing the database, and its use has been omitted from this analysis. Therefore, only the power meter and first wind speed sensor data are utilized in this analysis.

There are some situations where the wind power and speed measurement plots displayed missing values, so these must be dealt with before using the data from those respective months in further analyses. Otherwise, the observed values are in general agreement with the expected behaviour of the data for the months analysed in this phase.

Due to the descriptive nature of this analysis, the next two phases, data preparation and modelling are combined. The data preparation phase consists of following the steps outlined above in the data understanding section for specific months, in addition to computing the average wind speeds of the months in 2018 and searching for months with similar speeds in preceding years. Therefore, plots are generated during the data preparation phase. The final plots, average wind power and energy generated, represent the *model* in the scope of this analysis, hence the sections are combined into one descriptive section.

5 Data preparation and modelling

In this section, the data is prepared for analysis by checking quality of specific months, and searching for months that are suitable to compare. While this is, in principle, a part of the data understanding phase, it is a targeted understanding part needed for the descriptive modelling task, and therefore is kept separate from the general data understanding task. The final plots in this section are meant to represent the modelling phase, since the goal of this analysis is description. The analysis in this case is a comparison of the new turbine data with the old turbine data. Months need to first be analysed for the same average wind speed before being chosen for power comparison. To gain an understanding of the data quality, the same quality assessment described in the last section is done for all the relevant months in a systematic way. The history of the turbine operation and the availability of data is depicted in Figure 5-1.

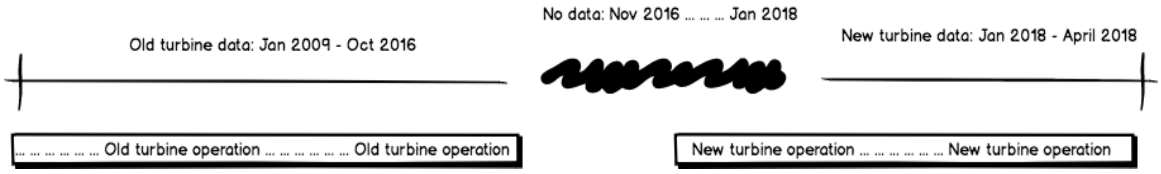

No data: Nov 2016 ... ... Jan 2018
Old turbine data: Jan 2009 - Oct 2016
New turbine data: Jan 2018 - April 2018
... ... Old turbine operation ... ... Old turbine operation
New turbine operation ... ... New turbine operation

Figure 5-1: The history of Ostfalia's HREP turbine operation 2009-2018

After running the preliminary quality check on the 2018 data, it was noticed that the minimum and maximum wind speeds were unusual. Upon closer visual inspection, it appeared that what was recorded does not represent natural wind speed. Rather, values stay constant for long periods of time, shown in Figure 5-2. The wind power was also plotted for a comparison to the wind speed, shown in Figure 5-3. The power values appear to have a similar behaviour compared to the old data, however they are more extreme in the sense that the peaks are higher though narrower and fewer in number. For comparison, the last four months of the old wind turbine are also analysed in the same fashion (Figure 5-4 and 5-5). The changes in the structure of the data appear to have happened after the installation of the new turbine. A check was also done for the period where there should be no data, and, indeed produced nonsensical values for some attributes.

Due to the unusual wind speed data in 2018, it is apparent that these data are not suitable for comparing the wind turbine performances. Nevertheless, this result can have some impactful insights for the HREP facility and is evaluated in the next section.

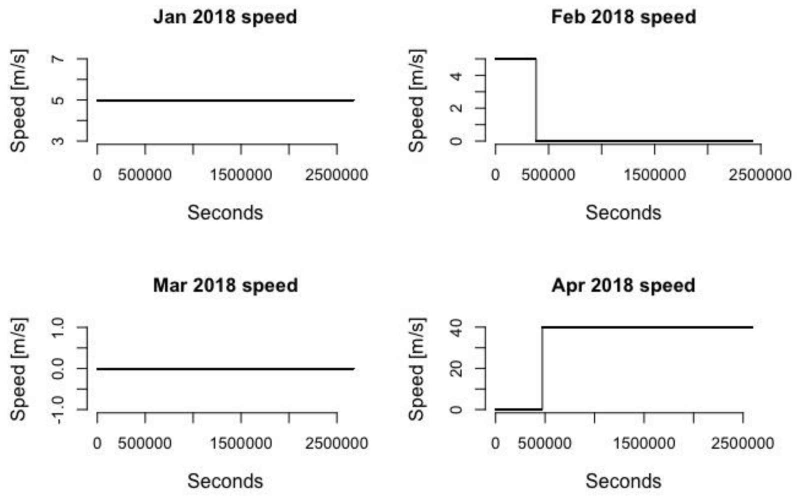


Figure 5-2: Visualization of wind speed in the new turbine (2018)

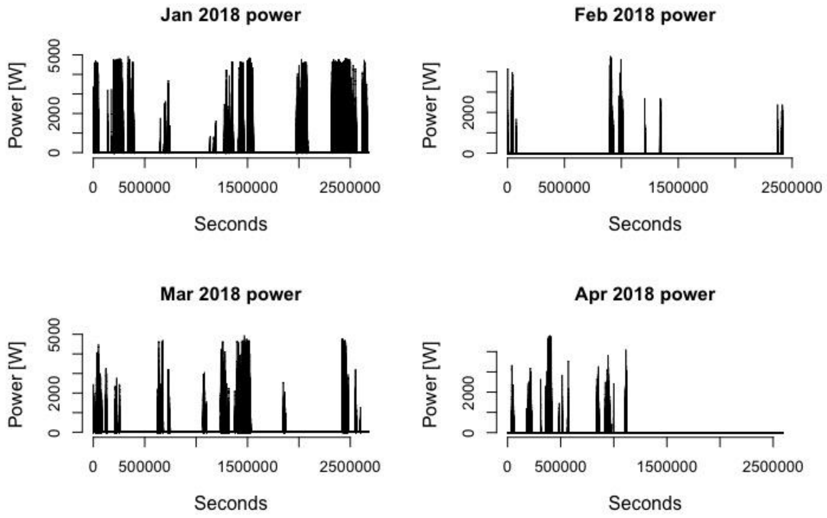


Figure 5-3: Visualization of wind power in the new turbine (2018)

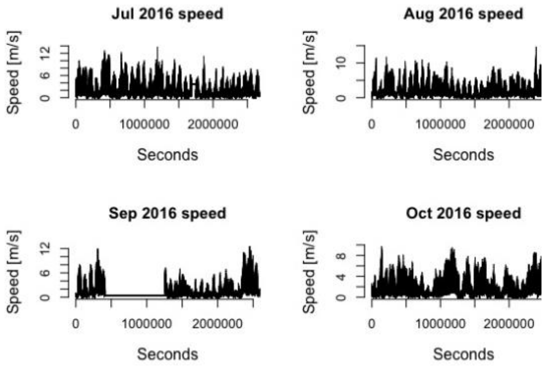


Figure 5-5: Wind speed with the old turbine (2016)

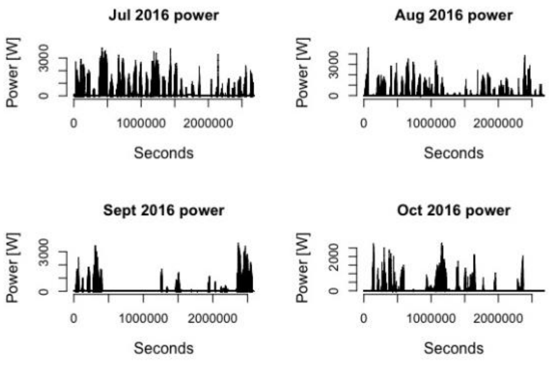


Figure 5-4: Wind power with the old turbine (2016)

6 Evaluation

The evaluation of descriptive data modelling is more abstract than the evaluation of other data mining tasks, which are associated with specific functions for evaluation [18]. In descriptive data modelling, the performance of the model has to do with its value to the data owners. Therefore, this section examines the potential value of the visualizations and the exploration with the CRISP-DM process to the facility.

The HREP facility has a need to understand the performance of its equipment, and this is especially important for the energy producing assets, although it is not limited to them; sensors and meters are just as important, because they are needed to acquire data in the first place and need to be taken care of.

This analysis started with a goal of illustrating the CRISP-DM process through a case study of comparing two turbines. However, through the process of data understanding and preparation, it was discovered that the data are not suitable for the task at the moment. Reasons for this are unclear, though a faulty sensor would not be an uncommon reason. The first step in the data preparation process was to realize the data quality and average wind speed in months when the new turbine was installed. In this process, it was discovered that the wind speed measurements in question do not follow the observed trends in the rest of the data, nor the recorded trends in literature, and are assumed to be inacurate.

This understanding provides knowledge to the HREP facility, albeit not the knowledge that was hoped for. At this stage, the choices are to re-evaluate the objective or close the project [18]. The exploratory analysis up to this point layed a solid foundation for continued research with the large amounts of data generated at the facility. The CRISP-DM process provided a systematic, and well documented workflow to come to this conclusion.

One of the objectives of this study was to report on the quality and accessibility of the data for potential future analyses. While the data availability and quality did not fulfil this analysis, it was noted during the study that there is a good opportunity for integrated tools for monitoring the data in more detail than what is currently possible with the existing tools. During this work, it was shown that making a direct connection to a MySQL database to access the HREP data through R is possible. The next section covers a possible solution to generate customized visualizations on the fly by coupling R with the local database.

7 Discussion and remarks

This section discusses the deployment phase of the CRISP-DM process, provides an outlook to the future and summarizes the work with concluding remarks.

7.1 Deployment: Using these results at the HREP

In the case of this analysis, deployment was not a continuous solution, rather a report with visualizations. Nevertheless, considerations for the future are included in this section, as well as remarks about value obtained from these results.

The methods, i.e. R code files, used to obtain the results of this work are included in Appendix A 1 (code regarding correlation is not included). They consist of functions written in the R language for importing and processing the data. These methods can be applied to wind power and speed data from different times by changing the import file names and variable names. Appendix A 1 also includes a function to connect to a MySQL database, which can be used to interact with data directly from the database and to access data not included in the MATLAB tool, such as wind direction data. These functions can be executed from a command line interface on a computer with R installed. These scripts can also be edited to evolve to different needs, such as outputting figures to files, or adding more advanced analytics.

The obtained results give a deep insight to the quality of the wind data at Ostfalia's HREP. These results are a starting point for building a better understanding of the facility through a data driven approach. The data is thought to be continuously generated in the same format, making these methods for basic analyses applicable to data from other times and as it is generated in the future. Things like syntactic accuracy and completeness of the data records were revealed to not have any major faults through this work, however the representativeness of the recorded values has been shown to vary. Hence, the visualizations applied here can (and should) be implemented with time periods not analysed in this work before using the data. Visualizations such as these can be used for diagnostic work, for example. It appears that there has been a change in the sensor behaviour after the installation of the new turbine, and this is a useful piece of information extracted from the data.

7.2 The future of renewable energy data

As infrastructure and energy demands continue to grow and develop, this development needs to be done sustainably to mitigate releasing excess carbon into the atmosphere or other types of pollution. A resource unique to this day and age is large amounts of data, which can be used to gain insights in developing sustainable energy solutions. Interest is growing in both of these areas and case studies exist, for example, for integrating big data into wave energy development [3].

While sensor, monitoring, and data storage technology advances give energy operators access to more and more data, there is a need to extract useful information from the data to drive innovative development in the energy sector. There is also a need to define what information is meaningful and how it should be presented to make real advances in nodes of energy networks, such as control centres [6] or building-level energy generation installations. Furthermore, cloud technology offers a promising alternative to storage and to computer processing power for managing and mining the large amounts of sensor and other data.

7.3 Summary

This work took on the challenge of gaining a better understanding of what to do with data generated from a hybrid renewable energy park (HREP) at Ostfalia University of Applied Sciences. While the desired analysis was not achieved, limitations in the data were discovered and documented, so that these data can be better utulized in the future. Also, the prospect of employing data mining techniques to a large database of a building-integrated hybrid renewable energy system to generate value was validated. The use of the Cross Industry Standard Process for Data Mining provided an indispensable framework for this project and is recommended as a basis for similar future projects.

References

[1] Destouni, Georgia; Frank, Harry: *Renewable Energy*. In: *Ambio* (2010) 39:1, p. 18 – 21

[2] Leventhal, Barry; *An introduction to data mining and other techniques for advanced analytics*. In: *Journal of Direct, Data and Digital Marketing Practice* (2010) 10:2, p. 137 – 153

[3] Amaro, Nuno; Pina João: *Big Data in Power Systems: Leveraging grid optimization and wave energy integration*. In: *International Conference on Engineering, Technology and Innovation (ICE/ITMC)* (2017), p. 1046 – 1054

[4] U.S. Energy Information Administration: *International Energy Outlook* (2011)

[5] Miniwatts Marketing Group, 2018: *Internet Usage Statistics* : *The Internet Big Picture*. https://www.internetworldstats.com/stats.htm

[6] Zhang, Pei; Li, Fangxing; Bhatt Navin: *Next-Generation Monitoring, Analysis, and Control for the Future Smart Control Centre*. In: *IEEE Transactions on Smart Grid* (2010) 1:2, p. 186 – 192

[7] Lou, Fang Lin; Hong, Y.: *Renewable Energy Systems*. Boca Raton: CRC Press Pub, 2017 http://proquest.safaribooksonline.com.elib.tamk.fi/book/electrical-engineering/power-systems/9781439891100

[8] Climate Science Investigations (CSI), 2016: *Global Wind Patterns*. http://www.ces.fau.edu/nasa/content/resources/global-wind-patterns.php

[9] Beccario, Cameron, 2018: *Earth*. https://earth.nullschool.net/

[10] Nabielek, Pia; Dumke, Hartmut; Weninger, Kurt: *Balanced renewable energy scenarios* : *a method for making spatial decisions despite insufficient data, illustrated by a case study of the Vorderland-Feldkirch Region, Vorarlberg, Austria*. In: *Energy, Sustainability and Society* (2018) 8:5

[11] Hordeski, Michael F*: Megatrends for Energy Efficiency and Renewable Energy*. Lilburn: The Fairmont Press Inc., 2010 https://ebookcentral.proquest.com/lib/tamperepoly-ebooks/detail.action?docID=3239037

[12] Gettelman, Andrew; Rood Richard: *Climate Change and Global Warming*. In: *Demystifying Climate Models* (2016), p. 23 – 35

[13] Bhandari, Binayak; Lee, Kyung-Tae; Lee, Gil-Yong; Cho, Young-Man; Ahn, Sung-Hoon: *Optimization of Hybrid Renewable Energy Power Systems: A Review*. In: *International Journal of Precision Engineering and Manufacturing-Green Technology* (2015) 2:1 p. 99 – 112

[14] Kirkham, Harold: *The Digital Revolution in Measurements*. In: *Innovative Smart Grid Technologies (ISGT)* (2013) p. 1 – 6

[15] Copernicus Programme, 2018: *CORINE Land Cover*. https://land.copernicus.eu/pan-european/corine-land-cover

[16] Google, n. d: *Google Maps Platform*. https://cloud.google.com/maps-platform/maps/

[17] Larose, Daniel: *Discovering Knowledge in Data* : *An Introduction to Data Mining*. Hobokin: Wiley-Interscience, 2005

[18] Berthold, Michael; Borgelt, C; Höppner, F; Klawonn, F: *Guide to Intelligent Data Analysis* : *How to Intelligently Make Sense of Real Data.* London: Springer-Verlag, 2010

[19] Big Data Ocean, n.d.: *The Big Data Ocean H2020 Project*. http://www.bigdataocean.eu/site/

[20] Bengtsson, Henrik; Jacobson, Andy; Riedy Jason: *R.matlab*. LGPL Software package. 2016-10-20

[21] Royal Academy of Engineering, n.d.: *Wind Turbine Calculations.* https://www.raeng.org.uk/publications/other/23-wind-turbine

[22] Baumann, Lars: *Improved System Models for Building-Integrated Hybrid Renewable Energy Systems with Advanced Storage* : *A Combined Experimental and Simulation Approach* (2015). PhD thesis, DeMontofrt University; Ostfalia University of Applied Sciences.

A 1 Appendix: R code files

R code files.

`1_import.R`:

```
if (!is.element("R.matlab", installed.packages()[,1]))
install.packages("R.matlab")
library(R.matlab)

setwd("…/2009")
data.1.2009 <- readMat("Mittelwerte_2009_01_1_SS.mat", verbose=TRUE)
setwd("…/RData")
save(data.1.2009, file="1.2009.RData")

setwd("…/2015")
data.1.2015 <- readMat("Mittelwerte_2015_05_1_SS.mat", verbose=TRUE)
setwd("…/RData")
save(data.1.2015, file="1.2015.RData")

setwd("…/2016")
data.1.2016 <- readMat("Mittelwerte_2016_01_1_SS.mat", verbose=TRUE)
setwd("…/RData")
save(data.1.2016, file="1.2016.RData")
```

`2_clean_form.R`:

```
# Clean and format data

# Take subset with columns power, speed and seconds
# Seconds, power and speed together in one data frame
process.wind <- function(data.in) {
  temp <- list(data.in$RegEnerg4, data.in$Wind)
  temp2 <- list(Seconds=temp[[1]][,1], Watts=temp[[1]][,2],
m_per_s=temp[[2]][,2])
  data.out <- as.data.frame(temp2, check.rows = TRUE)
  return(data.out)
}

# Check for completeness and NA values
comp.check <- function(data) {
  str(data) # check list lengths according to month
}

# Check for that data is numeric
syn.check <- function(data) {
  p <- as.numeric(data$Watts) # coerce to numbers
  s <- as.numeric(data$m_per_s) # coerce to numbers
  df <- data.frame(p, s)
  # Data are now numeric, check for any NA values:
  return(any(is.na(df)))
}

sem.check <- function(data) {
  min.p <- min(data$Watts)
  max.p <- max(data$Watts)
  min.s <- min(data$m_per_s)
  max.s <- max(data$m_per_s)
  print("min power | max power | min speed | max speed")
  sprintf("%.1f | %.1f | %.1f | %.1f", min.p, max.p, min.s, max.s)
}

q.check <- function(data) {
  cat("===============================\n\n")
  comp.check(data)
  print("NA values exist: ")
  print(syn.check(wdata))
  sem.check(data)
}
```

`3_plot.R`:

```
## Plots
# Functions to reduce the need to type styling into plot functions

setwd("…/OPlots")

# line plot (time series)
plot.ts <- function(x, y, filename) {
  xticks <- seq(1, 31, 1)
  plot(seq(1:length(x)), y, type="l",
       main="", xlab="", ylab="", axes=FALSE)
  axis(1)
  axis(2)
  mtext("Seconds", side=1, line=3)
}

histogram <- function(x) {
  hist(x, main="", xlab="")
}
```

`4_do.R`:

```
# Load data object
setwd("…/RData")
load(file="1.2009.RData")
load(file="1.2015.RData")
load(file="1.2016.RData")

# Take relevant subset of data
wind.1.2015 <- (process.wind(data.1.2015))
wind.1.2016 <- (process.wind(data.1.2016))

# Preliminary quality check
q.check(wind.1.2015)
q.check(wind.1.2016)

## 2015 time series

# Time series plot speed 2015
```

```
plot.ts(wind.1.2015$Seconds, wind.1.2015$m_per_s)
mtext("Speed [m/s]", side=2, line=3)
title("January 2015 wind speed time series")

# Time series plot power 2015
plot.ts(wind.1.2015$Seconds, wind.1.2015$Watts)
mtext("Power [W]", side=2, line=3)
title("January 2015 wind power time series")


## 2016 time series
par(mfrow=(c(1,2)))

# Time series plot speed 2016
plot.ts(wind.1.2016$Seconds, wind.1.2016$m_per_s)
mtext("Speed [m/s]", side=2, line=3)
title("January 2016 wind speed time series")

# Time series plot power 2016
plot.ts(wind.1.2016$Seconds, wind.1.2016$Watts)
mtext("Power [W]", side=2, line=3)
title("January 2016 wind power time series")

## Histograms
par(mfrow=(c(2,2)))

# 2015 speed
histogram(wind.1.2015$m_per_s)
title("Histogram of January 2015 speed")
mtext("Speed [m/s]", side=1, line=3)

# 2015 power
histogram(wind.1.2015$Watts)
title("Histogram of January 2015 power")
mtext("Power [W]", side=1, line=3)

# 2016 speed
histogram(wind.1.2016$m_per_s)
title("Histogram of January 2016 speed")
mtext("Speed [m/s]", side=1, line=3)

# 2016 power
histogram(wind.1.2016$Watts)
title("Histogram of January 2016 power")
mtext("Power [W]", side=1, line=3)

# Box plots 2015
par(mfrow=(c(1,2)))
boxplot(wind.1.2015$m_per_s, main="Box plot wind speed Jan 2015",
ylab="Speed [m/s]", cex=1)
boxplot(wind.1.2015$Watts, main="Box plot wind speed Jan 2015",
ylab="Power [W]", cex=1)
```

Modified script used in data preparation chapter.

`4a_do.R`:

```
#######################################################################
#####################
# Data preparation for monthly comparisons of wind speed and power
#######################################################################
#####################

# Months to analyze:
#  (Old turbine)
# 1-4.2018 (New turbine)

## Import statements:

if (!is.element("R.matlab", installed.packages()[,1]))
install.packages("R.matlab")
library(R.matlab)

## 2018
setwd("…/2018")
data.1.2018 <- readMat("Mittelwerte_2018_01_1_SS.mat", verbose=TRUE)
data.2.2018 <- readMat("Mittelwerte_2018_02_1_SS.mat", verbose=TRUE)
data.3.2018 <- readMat("Mittelwerte_2018_03_1_SS.mat", verbose=TRUE)
data.4.2018 <- readMat("Mittelwerte_2018_04_1_SS.mat", verbose=TRUE)
## 2016
setwd("…/2016")
data.7.2016 <- readMat("Mittelwerte_2016_07_1_SS.mat", verbose=TRUE)
data.8.2016 <- readMat("Mittelwerte_2016_08_1_SS.mat", verbose=TRUE)
data.9.2016 <- readMat("Mittelwerte_2016_09_1_SS.mat", verbose=TRUE)
data.10.2016 <- readMat("Mittelwerte_2016_10_1_SS.mat", verbose=TRUE)

## 2018
setwd("…/RData")
save(data.1.2018, file="1.2018.RData")
save(data.2.2018, file="2.2018.RData")
save(data.3.2018, file="3.2018.RData")
save(data.4.2018, file="4.2018.RData")
## 2016
save(data.7.2016, file="7.2016.RData")
save(data.8.2016, file="8.2016.RData")
save(data.9.2016, file="9.2016.RData")
save(data.10.2016, file="10.2016.RData")

## Extract attributes
## 2018
wind.1.2018 <- (process.wind(data.1.2018))
wind.2.2018 <- (process.wind(data.2.2018))
wind.3.2018 <- (process.wind(data.3.2018))
wind.4.2018 <- (process.wind(data.4.2018))
## 2016
wind.7.2016 <- (process.wind(data.7.2016))
wind.8.2016 <- (process.wind(data.8.2016))
wind.9.2016 <- (process.wind(data.9.2016))
wind.10.2016 <- (process.wind(data.10.2016))
```

```
## Quality check
## 2018
q.check(wind.1.2018)
q.check(wind.2.2018)
q.check(wind.3.2018)
q.check(wind.4.2018)

## Visual check
## 2018
par(mfrow=(c(2,2)))
# 1.2018
plot.ts(wind.1.2018$Seconds, wind.1.2018$m_per_s)
mtext("Speed [m/s]", side=2, line=3)
title("Jan 2018 speed")
# 2.2018
plot.ts(wind.2.2018$Seconds, wind.2.2018$m_per_s)
mtext("Speed [m/s]", side=2, line=3)
title("Feb 2018 speed")
# 3.2018
plot.ts(wind.3.2018$Seconds, wind.3.2018$m_per_s)
mtext("Speed [m/s]", side=2, line=3)
title("Mar 2018 speed")
# 4.2018
plot.ts(wind.4.2018$Seconds, wind.4.2018$m_per_s)
mtext("Speed [m/s]", side=2, line=3)
title("Apr 2018 speed")

# Check power
par(mfrow=(c(2,2)))
# 1.2018
plot.ts(wind.1.2018$Seconds, wind.1.2018$Watts)
mtext("Power [W]", side=2, line=3)
title("Jan 2018 power")
# 2.2018
plot.ts(wind.2.2018$Seconds, wind.2.2018$Watts)
mtext("Power [W]", side=2, line=3)
title("Feb 2018 power")
# 3.2018
plot.ts(wind.3.2018$Seconds, wind.3.2018$Watts)
mtext("Power [W]", side=2, line=3)
title("Mar 2018 power")
# 4.2018
plot.ts(wind.4.2018$Seconds, wind.4.2018$Watts)
mtext("Power [W]", side=2, line=3)
title("Apr 2018 power")

## 2016
# Speed
par(mfrow=(c(2,2)))
# 7.2016
plot.ts(wind.7.2016$Seconds, wind.7.2016$m_per_s)
mtext("Speed [m/s]", side=2, line=3)
title("Jul 2016 speed")
# 8.2016
plot.ts(wind.8.2016$Seconds, wind.8.2016$m_per_s)
mtext("Speed [m/s]", side=2, line=3)
title("Aug 2016 speed")
# 9.2016
```

```
plot.ts(wind.9.2016$Seconds, wind.9.2016$m_per_s)
mtext("Speed [m/s]", side=2, line=3)
title("Sep 2016 speed")
# 10.2016
plot.ts(wind.10.2016$Seconds, wind.10.2016$m_per_s)
mtext("Speed [m/s]", side=2, line=3)
title("Oct 2016 speed")

# Check power
par(mfrow=(c(2,2)))
# 7.2016
plot.ts(wind.7.2016$Seconds, wind.7.2016$Watts)
mtext("Power [W]", side=2, line=3)
title("Jul 2016 power")
# 8.2018
plot.ts(wind.8.2016$Seconds, wind.8.2016$Watts)
mtext("Power [W]", side=2, line=3)
title("Aug 2016 power")
# 9.2016
plot.ts(wind.9.2016$Seconds, wind.9.2016$Watts)
mtext("Power [W]", side=2, line=3)
title("Sept 2016 power")
# 10.2016
plot.ts(wind.10.2016$Seconds, wind.10.2016$Watts)
mtext("Power [W]", side=2, line=3)
title("Oct 2016 power")

# Av power 2018
cat(
  sprintf("Mean power 1.2018: %.2f Watts\n", mean(wind.1.2018$Watts)),
  sprintf("Mean power 2.2018: %.2f Watts\n", mean(wind.2.2018$Watts)),
  sprintf("Mean power 3.2018: %.2f Watts\n", mean(wind.3.2018$Watts)),
  sprintf("Mean power 4.2018: %.2f Watts\n", mean(wind.4.2018$Watts)))
```

Script to connect to a MySQL database.

`` `windDB.R`: ``

```
# Connect to MySQL database instance

install.packages("RMySQL")
# Load the library
library(RMySQL)

# Connect
db = dbConnect(MySQL(), user='user', password='password',
               dbname='windDB', host='localhost')

# db Operations
dbListTables(db)
dbListFields(db, '2016_01_windgeschwindigkeit')
new.table <- dbGetQuery(db, "SELECT * FROM
2016_01_windgeschwindigkeit")
```