BACHELOR

Condition based prognostics and diagnostics for wind turbines

Kenbeek, T.A.

Award date:
2016

Disclaimer
This document contains a student thesis (bachelor's or master's), as authored by a student at Eindhoven University of Technology. Student theses are made available in the TU/e repository upon obtaining the required degree. The grade received is not published on the document as presented in the repository. The required complexity or quality of research of student theses may vary by program, and the required minimum study period may vary in duration.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
Condition Based Prognostics and Diagnostics for Wind Turbines

Bachelor Thesis

T.A. Kenbeek
0721265

Supervisors:
dr. S. Kapodistria
dr. A. Di Bucchianico

Eindhoven, May 2016
Abstract

Condition based monitoring is widely present in wind turbine maintenance, as well as many other industries. Since wind turbines are complex mechanical systems, this requires extensive monitoring by qualified experts. In an era where data is easier and easier to generate and to store, we look for a fact-based statistical approach to this problem. This thesis observes a data set of a particular turbine and discusses various potential models which function as indicators of failure in wind turbines, as well as the methods used to create these models. We create estimates for temperature and power output levels, as well as control limits for these estimates. These limits are all dynamic, and are based on environmental variables, conditional variables, and maintenance and event logs. These models allow for an accurate and timely identification of failures of wind turbine parts.

Keywords: condition based monitoring; fault detection; statistical process control; predictive modelling.
Preface

This bachelor thesis is the result of my Bachelor Final Project at the department of Mathematics and Computer Science and is part of my Bachelor of Science study at the Eindhoven University of Technology.

First of all, I would like to thank Stella Kapodistria and Alessandro Di Bucchianico for their guidance and supervision throughout the project and for keeping their faith in me all this time. Without their feedback, help, and proof-reading, this thesis might not have made it past the first chapter. I would also like to thank the good people at [Oliveira] and [Delta] for sharing their data and giving me the opportunity to conduct this research.

Of course, I would also like to thank my mother and father, my three brothers and my sister for their unconditional love and support. Furthermore, my deepest thanks go out to my girlfriend Trudie, my best friends Bart, Laurens and Jet, as well as all my fraternity brothers from BALDR, for keeping me sane and happy.

Thomas Kenbeek
May 2016
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The photo of the wind park</td>
</tr>
<tr>
<td>1.2</td>
<td>The interior of the V47 nacelle</td>
</tr>
<tr>
<td>3.1</td>
<td>The environmental temperature, nacelle temperature, and oil temperature versus time</td>
</tr>
<tr>
<td>3.2</td>
<td>The gearbox temperature and bearing temperature</td>
</tr>
<tr>
<td>3.3</td>
<td>The main generator and starting generator temperature</td>
</tr>
<tr>
<td>3.4</td>
<td>Scatter plot of the temperature of the main generator for periods when it is disabled vs. the environmental temperature</td>
</tr>
<tr>
<td>3.5</td>
<td>Wind speed versus time</td>
</tr>
<tr>
<td>3.6</td>
<td>Power output versus rotor speed and generator speed</td>
</tr>
<tr>
<td>3.7</td>
<td>Power output. The lines indicating maintenance have been changed to black for legibility</td>
</tr>
<tr>
<td>3.8</td>
<td>Power output versus rotor speed</td>
</tr>
<tr>
<td>3.9</td>
<td>Power output versus wind speed, split based on RPM which we believe to correspond to the main generator (25.8), the starting generator (19−21) and generator unknown (other)</td>
</tr>
<tr>
<td>3.10</td>
<td>Three examples of vibration readings with unrealistically high values. The bottom right graph is an example of a measurement that seems to be calibrated wrong, but could still be useful</td>
</tr>
<tr>
<td>3.11</td>
<td>Examples of vibration readings that exhibit different patterns at different times. This is most likely due to re-calibration and unfortunately means the readings are not usable for this purpose</td>
</tr>
<tr>
<td>4.1</td>
<td>Shewhart chart of the residuals of the oil temperature, using an in-control period and 4 minute intervals</td>
</tr>
<tr>
<td>4.2</td>
<td>Scatter plot of oil temperature versus environmental temperature. The red line corresponds to the prediction interval of the regression model, the blue line corresponds to the control limits of the Shewhart chart</td>
</tr>
<tr>
<td>4.3</td>
<td>Correlation coefficient of oil temperature and environmental temperature as function of time</td>
</tr>
<tr>
<td>4.4</td>
<td>Correlation coefficient of nacelle temperature and environmental temperature as function of time. The green vertical line depicts the in-control moment with the highest correlation coefficient</td>
</tr>
<tr>
<td>4.5</td>
<td>Auto-correlation function for bearing temperature (left) and main generator temperature (right)</td>
</tr>
<tr>
<td>5.1</td>
<td>Autocorrelation of the residuals of the nacelle temperature</td>
</tr>
<tr>
<td>5.2</td>
<td>Shewhart control chart for the nacelle temperature when using an in-control period to calculate the regression model, and using 4 minute intervals</td>
</tr>
<tr>
<td>5.3</td>
<td>Confidence interval for the nacelle temperature when using an in-control period to calculate the regression model, and using 4 minute intervals</td>
</tr>
<tr>
<td>5.4</td>
<td>Nacelle temperature versus time. The points which are identified as warnings are color-characterized in red</td>
</tr>
<tr>
<td>5.5</td>
<td>Autocorrelation plot for oil temperature</td>
</tr>
<tr>
<td>5.6</td>
<td>Shewhart chart for the residuals of the oil temperature, using 4 minute intervals</td>
</tr>
<tr>
<td>5.7</td>
<td>Scatter plot of oil temperature and environmental temperature, with lines representing the point estimates (green) and the prediction interval (red). The points which are classified as warnings according to the prediction interval are color-characterized in red</td>
</tr>
</tbody>
</table>
5.8 Oil temperature versus time. The points which are classified as warnings according to the Shewhart chart are color-characterized in red. ........................................... 43
5.9 Autocorrelation for the bearing temperature, using the entire data set in the top picture and the reduced data set (4-hour intervals) in the bottom picture. .......... 45
5.10 Shewhart chart for the residuals of the bearing temperature ..................... 46
5.11 Bearing temperature. Warnings are color-characterized in red. ................. 47
5.12 Autocorrelation for the gearbox temperature, using the entire data set in the top picture and the reduced data set (4-hour intervals) in the bottom picture. .... 48
5.13 Shewhart chart for the residuals of the gearbox temperature, using 4 minute intervals 49
5.14 Gearbox temperature. Warnings are indicated in red. ............................. 50
5.15 Autocorrelation plot for the main generator temperature, using the entire data set in the top plot, and the reduced (4h-intervals) in the bottom plot ............... 51
5.16 Shewhart chart for the residuals of the main generator temperature .......... 52
5.17 Main generator temperature. Warnings are color-characterized in red. ........ 53
5.18 Auto-correlation plot for the starting generator temperature, using 4 minute and 4 hour intervals ................................................................. 54
5.19 Shewhart chart for the residuals of the starting generator temperature ........ 55
5.20 Starting generator temperature. Warnings are color-characterized in red. .... 56
5.21 Theoretical power curve for the Vestas V47 ........................................ 57
5.22 Autocorrelation for the power output of the main generator, using the entire data set in the top picture and the reduced data set (4-hour intervals) in the bottom picture. ...................................................... 58
5.23 Shewhart chart for the residuals of the power output, using data for the main generator only ................................................................. 59
5.24 Power output for the main generator. Warnings are color-characterized in red. 60
5.25 Autocorrelation for the power output of the starting generator temperature, using the entire data set in the top picture and the reduced data set (4-hour intervals) in the bottom picture. ...................................................... 61
5.26 Shewhart chart for the residuals of the power output, using data for the starting generator only ................................................................. 62
5.27 Power output for the main generator. Warnings are color-characterized in red. 63
# List of Tables

1.1 The different parts of the Vestas V47-turbine ........................................ 8  
2.1 Wind turbine assemblies’ reliability field data ........................................... 9  
2.2 General set of wind turbine assemblies and main parts ............................... 10  
2.3 Root causes of failure modes ....................................................................... 10  
2.4 Major detection methods of the failure modes .............................................. 10  
3.1 The available variables in the data set excluding subdivision of the vibration readings into bands ................................................................. 12  
3.2 Correlation coefficients for nacelle temperature, oil temperature and environmental temperature ................................................................. 14  
3.3 List of vibration variables that behave in-control for the entire data period .... 16  
5.1 Comparison of $R^2$ and BIC values for the nacelle temperature regression models using 4 minute or 4 hour intervals ......................................................... 36  
5.2 Regression model of the nacelle temperature, 4 hour intervals ..................... 37  
5.3 Comparison of $R^2$ and BIC values for the oil temperature .......................... 39  
5.4 Regression model of the oil temperature, 4 hour intervals ............................ 40  
5.5 Comparison of $R^2$ and BIC values for the bearing model, using 4 minute or 4 hour intervals ................................................................. 44  
5.6 Regression model for the bearing temperature, four hour intervals ............... 45  
5.7 Comparison of $R^2$ and BIC values for the gearbox temperature, using 4 minute and 4 hour intervals ................................................................. 48  
5.8 Regression model for the gearbox temperature, 4 hour intervals ................. 48  
5.9 Comparison of $R^2$ and BIC values for the main generator temperature, using 4 minute and 4 hour intervals ................................................................. 51  
5.10 Regression model for the temperature of the main generator ...................... 51  
5.11 Comparison of $R^2$ and BIC values for the starting generator temperature, using 4 minute and 4 hour intervals ................................................................. 53  
5.12 Regression model for the temperature of the starting generator, four minute intervals ................................................................. 53  
5.13 Comparison of $R^2$ and BIC values for the power output of the main generator, using 4 minute and 4 hour intervals ................................................................. 58  
5.14 Regression model of the power output, using data from the main generator and 4 minute intervals ................................................................. 59  
5.15 Comparison of $R^2$ and BIC values for the power output of the main generator, using 4 minute and 4 hour intervals ................................................................. 60  
5.16 Regression model of the power output, using data from the starting generator and 4 minute intervals ................................................................. 60
Chapter 1

Problem statement

In this manuscript we analyze data provided to us by Oliveira for the OLAZ-1 wind turbine from the OLAZ wind park. The OLAZ wind park is located near Borssele in the province of Zeeland and is owned by Delta. The objective of the manuscript is the creation of models which permit the scheduling of predictive maintenance of wind turbines by predicting failures in a timely manner. To this purpose, we analyse a large amount of data, consisting of condition data, operational data and environmental data, together with maintenance and event logs, that have been collected over a period of almost two years. Our models use all provided data. In particular, we consider a baseline period and use the corresponding data to produce the models. Thereafter, we validate the models and get timely and accurate warnings on the imminent failures of the turbine components.

More concretely, our models can be distinguished in two categories: the first category consists of various modes constructed based on the various temperature measurements for specific components. The second category consists of a single model which represents an overall health index based on the power output.

Before the development of the aforementioned models, we need to better understand the structure of the data at hand and check for the underlying quality of the information. To this purpose, we first give some basic background information on the structure of the wind turbines. We then have a closer look at the data, and attempt to understand how the various parts behave. Subsequently, we construct the predictive models, based on temperature and power output. Finally, we discuss the results and draw conclusions.

1.1 Wind turbine description

The data was collected from the OLAZ-1 turbine. OLAZ-1 is a wind turbine of type Vestas-V47 and was built in 2000. The height of the turbine is 65 m, the rotor diameter is 47 m and the rotor sweep is 17.35 m². The typical power output for the Vestas V-47 turbine is 600 kW. It joins the grid connection at a wind speed of 4 m s⁻¹ with a rated actual power output (typically achieved) at a wind speed of 15 m s⁻¹, and it is disconnected at a wind speed of 25 m s⁻¹. Furthermore, the Vestas V-47 turbine is designed to function up to a maximum wind speed of 59.5 m s⁻¹.

Typically, the Vestas V-47 turbine is equipped with a single generator. However, the OLAZ-1 is equipped with two generators. The second generator is smaller and is only used when the wind speed is very low (less than 7 m s⁻¹). The main benefits of a second generator are a lower sound level and an increased power production at low wind speeds.

The turbine consists of several parts. These parts are listed in Table 1.1 and a visualization of the OLAZ-1 turbine is provided in Figure 1.2. On December 2014, the OLAZ-1 turbine suffered a mechanical failure of the main generator. While the domain experts suspected the problem, the identification of the root cause and the extend of the problem became apparent only after the failure. Our objective is to check whether the data could have assisted in identifying a change in the pattern and whether we could only from the data predict the failures well in advance. In this sense, one of the key objectives of this manuscript is to develop predictive models based on data so as to predict imminent failures, and identify the causal component.

Of course, since the objective is to perform predictive maintenance, our models should predict the imminent failures with sufficient accuracy while permitting us time-wise to identify the cause of the failure and repair the problem. Finally, it is worth mentioning that after the failure of the
generator, several other components such as the bearing and gearbox exhibited stress and failed (or were preventive replaced) soon after the generator failure.

<table>
<thead>
<tr>
<th>1. Blade</th>
<th>11. Service crane</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Main shaft</td>
<td>14. Tower</td>
</tr>
<tr>
<td>5. Secondary generator</td>
<td>15. Yaw control</td>
</tr>
<tr>
<td>6. Gearbox</td>
<td>16. Gear tie rod</td>
</tr>
<tr>
<td>7. Disc break</td>
<td>17. Yaw ring</td>
</tr>
<tr>
<td>8. Oil cooler</td>
<td>18. Yaw gears</td>
</tr>
<tr>
<td>9. Cardan shaft</td>
<td>19. VMP top control unit</td>
</tr>
<tr>
<td>10. Primary generator</td>
<td>20. Hydraulic unit</td>
</tr>
</tbody>
</table>

Table 1.1: The different parts of the Vestas V47-turbine
Chapter 2

Reliability and FMEA for wind turbines

Many studies have been conducted to evaluate the reliability of wind turbines as the integrated part of the grid, see, e.g., Hu et al. (2009), Kahrobaee and Asgarpoor (2010), Lingling and Yang (2010). Some of these studies have addressed the individual wind turbine reliability modeling, and investigated the major factors contributing to the total failure of the turbine, see, e.g., Spinato et al. (2009), Guo et al. (2009). Reliability data about wind turbine assemblies have become available in recent years from surveys, see, e.g., Tavner et al. (2007), Spinato et al. (2009). Table 2.1 shows a typical comparison between reliability field data of a small wind turbine, 300 kW, and a 1 MW wind turbine main assembly failure rates based on Spinato et al. (2009).

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Failure rate of LKW WTs (Failure per turbine per year)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>300 kW WT</td>
</tr>
<tr>
<td>Generator</td>
<td>0.059</td>
</tr>
<tr>
<td>Brake</td>
<td>0.029</td>
</tr>
<tr>
<td>Hydraulics</td>
<td>0.039</td>
</tr>
<tr>
<td>Yaw system</td>
<td>0.079</td>
</tr>
<tr>
<td>Sensors</td>
<td>0.037</td>
</tr>
<tr>
<td>Pitch system</td>
<td>0.034</td>
</tr>
<tr>
<td>Blade</td>
<td>0.078</td>
</tr>
<tr>
<td>Gearbox</td>
<td>0.079</td>
</tr>
<tr>
<td>Shaft/bearings</td>
<td>0.002</td>
</tr>
</tbody>
</table>

Table 2.1: Wind turbine assemblies’ reliability field data

Besides reliability analysis, failure mode and effects analysis (FMEA) is performed in order to determine several key potential failures in the system through the comparison of some predefined factors. As a result, such an analysis helps increase the availability of that system, see, e.g., Stamatis (2003), Mikulak et al. (2008). This process has been used on almost any equipment from cars to space shuttles, and as of the last decade wind turbines have been briefly studied, see, e.g., Klein and Lali (1990), Andrawus et al. (2006), Tavner et al. (2010).

2.1 Failure modes

The failure occurs when a device no longer operates the way intended. There are numerous failure modes that can be defined for a complicated assembly like a wind turbine. These failure modes can cause partial or complete loss of power generation. Mainly, the key failure modes, which cause complete loss of power generation, are malfunction and major damage of the main parts of the turbine. Other failure modes are less significant and may be surface damage and cracks, oil leakage, loose connection, etc. However, if they are not taken care of, minor failure modes can initiate major failures as well, see for more details Table 2.2 based on Kahrobaee and Asgarpoor (2011). Evidently, each one of the failure modes has a root cause, and the probability of that failure mode is directly related to the probability of its root cause. Table 2.3, based on Kahrobaee and Asgarpoor (2011), provides different categories for these causes. Human error in this table, refers to the errors occurring during operation or maintenance.
### Sub-assemblies
- **Structure**: Nacelle, Tower, Foundation
- **Rotor**: Blades, Hub, Air brake
- **Mechanical Brake**: Brake disk, Spring, Motor
- **Main shaft**: Shaft, Bearings, Couplings
- **Gearbox**: Toothed gear wheels, Pump, Oil heater/cooler, Hoses
- **Generator**: Shaft, Bearings, Rotor, Stator, Coil
- **Yaw system**: Yaw drive, Yaw motor
- **Converter**: Power electronic switch, cable, DC bus
- **Hydraulics**: Pistons, Cylinders, Hoses
- **Electrical System**: Soft starter, Capacitor bank, Transformer, Cable, Switch gear
- **Pitch System**: Pitch motor, Gears
- **Control system**: Sensors, Anemometer, Communication parts, Processor, Relays

### Table 2.2: General set of wind turbine assemblies and main parts

<table>
<thead>
<tr>
<th>Weather</th>
<th>Mechanical</th>
<th>Electrical</th>
<th>Wear</th>
</tr>
</thead>
<tbody>
<tr>
<td>High wind</td>
<td>Manufacturing and material defect</td>
<td>Grid fault</td>
<td>Aging</td>
</tr>
<tr>
<td>Icing</td>
<td>Human error</td>
<td>Overload</td>
<td>Corrosion</td>
</tr>
<tr>
<td>Lightning</td>
<td>External damage</td>
<td>Human error</td>
<td></td>
</tr>
<tr>
<td></td>
<td>External damage</td>
<td>Software failure</td>
<td></td>
</tr>
</tbody>
</table>

### Table 2.3: Root causes of failure modes

#### 2.2 Failure detection

There are a variety of ways to detect the probable failure modes as categorized in Table 2.4 based on Kahrobaee and Asgarpoor (2011). The common ways are through inspection or while the turbine is being maintained. However, the fastest and the most reliable method is condition monitoring which can increase the availability of the wind turbine considerably by using online systems. With condition monitoring, the probability of not detecting the failure decreases to the failure probability of the human error or the monitoring system itself. The objective of this manuscript is to suggest statistical approaches that can be used to analyze the data so as to detect imminent failures.

<table>
<thead>
<tr>
<th>Inspection</th>
<th>Condition Monitoring</th>
<th>Maintenance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual</td>
<td>Vibration analysis</td>
<td>Time-Based</td>
</tr>
<tr>
<td>Olfactive</td>
<td>Oil analysis</td>
<td>Condition-Based</td>
</tr>
<tr>
<td>Auditive</td>
<td>Infrared thermography</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ultrasonic</td>
<td></td>
</tr>
</tbody>
</table>

### Table 2.4: Major detection methods of the failure modes
Chapter 3

Data description

For the development of the statistical approaches with the objective of detecting imminent failures, we use a data set containing 326160 observations on 110 variables. The data were collected in the period from 19/06/2013 18:32:00 to 18/03/2015 23:56:00. The names of the variables are listed in Table 3.1. In addition to the data set, we use the maintenance and event logs for the period of interest. In particular, all information we use for the development of the statistical approaches can be clustered into three categories:

1. **Environmental variables**: These describe the environmental factors. In our data set, the environmental variables are the wind speed and the environmental temperature.

2. **Conditional variables**: These variables describe the state of several parts of the turbine. The conditional variables can be further categorized into
   - (a) **Speed**: This includes rotor speed and generator speed. Note that generator speed does not distinguish between the main generator and the starting generator.
   - (b) **Temperatures**: This includes readings of the bearing, the gearbox, the two generators, the nacelle and the oil temperatures.
   - (c) **Vibrations**: These include 29 different vibration measurements on 11 different components. As unit of measurement, acceleration is used.
   - (d) **Operational readings**: This includes the power output, the operating state, the pitch angle, and the yaw.

3. **Maintenance and event logs**: Apart from the data set, we were also provided with access to the processed maintenance logs of the past two years. Any time any maintenance is performed on the turbine, the date and time is logged, as well as a short description of the work that was done, and if applicable the corresponding result, e.g. “Ultra sonic scanning of blade bolds”.

We have to note that the overall vibration readings are separated into 9 bands according to the frequency levels. However, for reasons of simplicity and practicality, we decided as a first step to only use and mention the overall readings for our analysis. In the future, we plan to create models that identify the root cause of the imminent failure and for such an endeavor the separation of vibration readings into the various bands will be of critical value.

For the purpose of developing failure prediction models, we only use data collected during instances in which the turbine is running. To this end, we use the variable “operating state”. This variable takes 4 values: 0 to indicate emergency, 1 to indicate stop, 2 to indicate pause and 3 to indicate run. We restrict our analysis to the cases the variable operating state takes value 3. This leaves us with 138169 measurements.

Although the available data span till March 3, 2015, in some cases, we restrict our analysis to the data collected before December 15, 2014. This is the date that the main generator failed. While after this date the turbine continued to operate, it is only operating based on the starting generator. This creates some irregular patterns to the data collected after the time of failure of the main generator.

### 3.1 Data visualization

As a first step, so as to get a preliminary understanding of the data we present in this section plots of the individual variables versus time, as well as some scatter plots in order to visually
Table 3.1: The available variables in the data set excluding subdivision of the vibration readings into bands

identify patterns and relations between variables. More concretely, we want to visualize the data for three reasons:

- Visually check the quality of the data. This means identify any anomalies, such as huge gaps or unrealistic values.
- Identify patterns related to the generator failure. If we see certain variables behaving very differently after the generator failure, that could be of interest in the creation of algorithms that predict this failure.
- Identify interesting relationships between variables to be used for our predictive models.

3.1.1 Temperature

The most interesting and informative variables are those related to the temperature. In the plots that follow, when we plot a certain variable over time, we add to the plot vertical lines that signify important maintenance events. The red lines indicate

- Large maintenance on the entire turbine, November 16, 2013.
- A complete disassemble of the main generator, March 28, 2014.
- The deactivation of the main generator, December 15, 2014.

The blue lines indicate

- The removal of several sensors and cables, November 18, 2014.
- The placement of sensors on the generator and the shaft, March 23, 2014.

The green line indicates the end of the in-control period that we use throughout the report. How this exact period is determined, is discussed in detail in section 4.2.

First, we plot in Figure 3 the environmental, the nacelle and the oil temperature versus time. For these three figures we note that the temperature exhibits seasonal behavior. Moreover, from the visual observation of the three plots in Figure 3, we can conjecture that the environmental, the nacelle and the oil temperatures all behave very similarly. This can be statistically confirmed by considering the correlation coefficient. In particular, let $\rho_{e,n}$ denote the correlation coefficient
Figure 3.1: The environmental temperature, nacelle temperature, and oil temperature versus time

between the environmental (ambient) temperature and the nacelle temperature for the entire
time period, and similarly for the other variables. The results are shown in Table 3.2.

Next, in Figure 3.2, the temperatures of the gearbox temperature and the bearing temperature
are plotted. In Figure 3.3, the temperature of the two generators are plotted against time.

The gearbox and bearing temperature behave very similarly. This is because of their physical
connection that causes a direct heat transfer.

If we look at the temperature of either generator, we observe that the minimum generator temperature seems to behave similarly to the environmental temperature, while the maximum temperature does not seem to be influenced by the environmental temperature. This can be explained by the fact that the generators only generate heat when they are in use. When they are not in use, the only source of heat comes from the nacelle cooling system, which is almost identical to the environmental temperature, see Figure 3.4.

Our conjecture, that the minimum temperature matches the environmental temperature almost perfectly, can be easily proven by investigating the temperature of the main generator during times when it was disabled. We expect to see a high correlation with the environmental temperature. Indeed the correlation coefficient between the environmental temperature and the generator temperature after November 16, 2014 is $\rho_{g,e} = 0.9035485$, which indicates an almost perfect linear correlation.

Other variables of interest include the power output, the wind speed, the rotor speed and the generator speed. We see a sharp drop in power output around the end of 2014. This is explained by the failure of the main generator. The maximum output of the starting generator is lower than the output of the main generator at higher wind speeds. We can visually identify low, moderate and high wind speeds through color coding, and plot the power output versus time. The power seems to exhibit one pattern of behavior up to December 2014. We know from the maintenance logs that the main generator was deactivated around December 2014, so this is why

<table>
<thead>
<tr>
<th>correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{e,o} = 0.8941807$</td>
</tr>
<tr>
<td>$\rho_{n,o} = 0.961099$</td>
</tr>
</tbody>
</table>

Table 3.2: Correlation coefficients for nacelle temperature, oil temperature and environmental temperature

Figure 3.2: The gearbox temperature and bearing temperature
we see the sharp drop in power output.

The rotor and generator speed behave practically identically. This makes sense, since the rotor is attached to the generator. One directly propels the other. This can be observed similarly as before by directly calculating the correlation coefficient. In this case, we observe that 

$$\rho_{r,g} = 0.9997668.$$  

Furthermore, we notice a large concentration of values around 20 RPM and 25 RPM, and a big gap in between those two values. These large concentrations are created because each generator only works at one specific speed. The starting generator generates power around 20 RPM and the main generator generates power at 25 RPM.

Next, it is useful to look at some scatter plots, to visually identify relationships between the various variables. First, we take a closer look at the relationship between the rotor speed and the power output.

In Figure 3.8, the scatter plot of the rotor speed versus the power output is depicted. This plot should be read as follows: every paired observation of the rotor speed with the corresponding power output is depicted on the plot as a dot and is color-characterized based on the wind speed. In particular, if the wind speed is below 3.5 m/s, it is color-coded green in the plot. If the wind speed is between 3.5 m/s and 14 m/s it is color-coded blue and if the wind speed exceeds 14 m/s, it is color-coded red. Note that the depicted plot resembles the letter “F”, where the lower vertical line corresponds to power production achieved by the starting generator, while the upper vertical line corresponds to power production achieved by the main generator. This is in accordance with the fact that the two generators work at two distinct speeds. This will later prove to be very useful, as there is no variable directly indicating which of the two generators is in use. Based on the rotor speed, we can determine which one of the two generators is in use. From this point onward we formulate 3 distinct cases and corresponding assumptions for the generator in use:

i. When the rotor speed exceeds 25.8 rotations per minute (RPM), then we may assume that the main generator is in use.

ii. When the rotor speed exceeds 19 RPM but does not exceed 21 RPM, then we may assume that the starting generator is in use.

iii. When the rotor speed is below 19 RPM, since in these cases there is also no power output, it seems likely that neither generator is in use at those times.

In the rest of the manuscript, when we need to identify whether the main or the starting generator is in use, we will apply rules i) and ii), respectively. To this end, we create a variable, say GenInUse, that can take values 0 (for unknown generator in use), 1 or 2 (for the main and the starting generator in use, respectively).

Applying the above rules, we create the scatter plots of the power output of the main generator and the starting generator versus the wind speed. These are depicted in Figure 3.9. We see that all the points with a rotor speed of less than 19 RPM correspond very neatly to either a power output of 0 or a negative power output (consumption of power). This reinforces our hypothesis that a rotor speed below 19 RPM means that neither generator is in use. For the power output corresponding to either the main generator or the starting generator, the plots seem to indicate a cubic relation. This relation is also a theoretical result based on physical model connecting the wind speed with the power output, that will be discussed further in Section 5.2.

## 3.2 Vibration data

One of the main issues that we encountered, was that the quality of some vibration readings seemed questionable.

### 3.2.1 Reasonableness of vibration quality

The vibration variables contain some very unrealistic values. The vibrations are measured in G, with 1G = 9.81 m s\(^{-2}\), the average gravitational acceleration. Values up to about 1G would seem reasonable. However, several readings provide us with spike values of several hundreds or even thousands of G. This typically occurs for a period of about a month and does not occur outside this period. This may have to do with loose cables of the sensor or other mechanical sensor issues. In some other cases, unrealistic values occur during the entire measurement period. This may simply have to do with a wrong sensor calibration. In these cases, we can assume that the results are still useful, because this would mean the results are a scalar multiple of the true value, and that would not influence the analysis. Some examples are plotted in Figure 3.10.
3.2.2 Consistency of vibration quality

We see that some vibration variables exhibit different patterns at different times, for no apparent reason. This may have to do with a re-calibration of the sensor. Some examples are plotted in Figure 3.11.

There are several problems with the appearance of unrealistic values and patterns. First, it makes it hard to visually draw any conclusions from the vibration data, since the plots are scaled up to the largest value, so outliers make the smaller values more difficult to distinguish, and smaller readings become indistinguishable. E.g. if a plot’s y-axis is scaled from 0 to 100, the difference between a reading of 0.35 or 0.45 is very hard to tell, even though it might be significant for our analysis. Second, if the sensor has provided us with obviously erroneous data, how can we be sure that the values that seem “normal” are correct? We cannot. This, unfortunately means that a large amount of data is unusable for analysis. A complete list of vibration variables that behave according to our expectations is in Table 3.3

<table>
<thead>
<tr>
<th>GbxMSA10KHzOvr</th>
<th>GbxMSA1KHzOvr</th>
</tr>
</thead>
<tbody>
<tr>
<td>GbxMSD500HzOvr</td>
<td>GbxOSA10KHzOvr</td>
</tr>
<tr>
<td>GbxOSA2KHzOvr</td>
<td>GbxOSD1KHzOvr</td>
</tr>
<tr>
<td>GenDED1KHzOvr</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3: List of vibration variables that behave in-control for the entire data period of the vibration data seems to be of no use for our purposes.

3.3 Conclusions

In this chapter we were able to draw some important conclusions:

- There is a very strong correlation between the environmental temperature and the nacelle temperature. To a lesser extent, this is also true for the oil temperature.

- In order to model the generator temperature, gearbox temperature and bearing temperature, we need additional variables besides temperature.

- Unfortunately, only a small number of vibration variables seem to be useful for analysis.

- We can identify which generator is being used by looking at the rotor speed. A rotor speed around 20 RPM indicates that the starting generator is in use, while a rotor speed of 25 RPM indicates that the main generator is in use.

- There seems to be a cubic relation between wind speed and power output. This will be discussed further in section 5.2.
Figure 3.3: The main generator and starting generator temperature
Figure 3.4: Scatter plot of the temperature of the main generator for periods when it is disabled vs. the environmental temperature
Figure 3.5: Wind speed versus time
Figure 3.6: Power output versus rotor speed and generator speed
Figure 3.7: Power output. The lines indicating maintenance have been changed to black for legibility.
Figure 3.8: Power output versus rotor speed
Figure 3.9: Power output versus wind speed, split based on RPM which we believe to correspond to the main generator (25.8), the starting generator (19 - 21) and generator unknown (other).
Figure 3.10: Three examples of vibration readings with unrealistically high values. The bottom right graph is an example of a measurement that seems to be calibrated wrong, but could still be useful.
Figure 3.11: Examples of vibration readings that exhibit different patterns at different times. This is most likely due to re-calibration and unfortunately means the readings are not usable for this purpose.
Chapter 4

Statistical process control

From the problem statement, recall that the purpose of this report is to produce models which will predict imminent failure of any of the wind turbine parts or the turbine as a whole. The way to do this, is to detect deviations from what we consider to be “normal behavior” in these models. To identify what qualifies as normal behavior and to determine what deviates from normal behavior, we will employ techniques from a group of methods known as statistical process control (SPC).

SPC is a method of quality control which uses statistical methods. It is applied in order to monitor and control a process by looking at deviations from normal behavior. SPC originated in the manufacturing industry, as a way to monitor the quality and consistency of the manufactured items. It goes back to Walter Shewhart, who introduced the concept in 1924, see Shewhart (1924)

The standard setting for control charts is to collect small groups of observations at distinct time points. These groups usually occur in a natural manner. They could consist, for example, of items that were manufactured simultaneously. These groups are called rational subgroups. Typically, the mean and/or the standard deviation of each group is assessed for “normal” behavior. In SPC-terminology, normal behavior is called in-control, while deviations from normal behavior are called out-of-control. Control is actually a misnomer in this case. It is not about control in the engineering sense of the word, like in feedback control. The processes are actually about monitoring. The word “control” is still used for historical reasons.

The main idea behind SPC is that there is variation in every process, but that this variation can be described as being one of two types: in every process, there is intrinsic variation. This is variation that always occurs due to the nature of the process, which will never be 100 % deterministic. Shewhart called this variation due to common causes. This type of variation always occurs, even in an in-control situation. Secondly, there is variation due to special causes. This is the type of variation that defines an out-of-control situation, and the purpose of SPC is exactly to detect variation of this type. The reason rational subgroups are used is to get an unbiased estimator of the process variance. This is then compared to the level of variation which one expects in an in-control situation (the variation due to common causes). The simplest way to make these ideas operational is through the so-called Shewhart control charts. Such a chart is a way to visualize the variation of a process and to compare it to what one expects due to common causes.

The underlying statistical model is assumed to be independent identically distributed (iid) normal observations with the same mean in each rational subgroup. Should anything change in the process due to special causes, then this should manifest itself as changes between the rational subgroups.

The control limits are usually set at $\bar{x} \pm 3 \frac{\hat{\sigma}}{\sqrt{n}}$, where $\bar{x}$ and $\hat{\sigma}$ are determined from historical data. This data has to be from a period which we know to be in-control, or a start-up period. This period is usually called phase I. Other choices exist, such as the range estimator for the standard deviation. This is turned into an unbiased estimator by using a multiplicative constant.

The actual monitoring is called phase II. The control chart gives a signal when the mean of a rational subgroup is outside the control limit. For an up to date review of phase I and phase II, see Hawkins et al. (2003).

It is important that the control chart signals as quickly as possible when an out-of-control situation has occurred, while at the same time the number of false alarms should be minimal. The proper way to quantify these performance characteristics is to consider the so-called run length distributions. The in-control run length of a control chart is the time to a signal (alarm) assuming that the process stays in-control. Note that this is a random variable. Shewhart
control charts have no memory due to the independence of the observations, and it is easy to see that the run length has a geometric distribution. A commonly used quantity is the in-control average run length (ARL), which is simply the expected number of observations until the first signal appears in an in-control situation. Since the expected value of a geometric distribution is \( E[\text{geom}] = \frac{1}{p} \), this gives us an ARL of 370. Note however that ARL may be misleading since run length distributions are non-negative, and therefore usually skewed. For an in-depth treatment of this topic, see Frisén (1992), Frisén (2007).

Similarly, there is an out-of-control run length, associated to the time it needs to detect an out-of-control situation. For this, one needs to specify an out-of-control situation. Common out-of-control situations are the persistent jump of the mean, which typically occurs when a piece of equipment breaks, and a gradual shift of mean, which typically occurs when a piece of equipment wears. For a more extensive discussion, see Di Bucchianico and Van den Heuvel (2015)

Apart from Shewhart charts, there are more advanced ways to monitor processes, like the exponentially weighted moving average (EWMA) and the cumulative sum (CUSUM). These are discussed in Montgomery (2005). For our purposes, Shewhart charts suffice and are easier to implement.

Not everything defined in the standard setting above is applicable to our situation. For a start, we are not dealing with independent variables. The power output, for instance, depends heavily on the wind speed and other factors. However, we can still determine control limits based on an in-control period. To achieve this, we develop regression models and monitor the (non-standardized) residuals. This is similar to what is called tool wear charts. Not to be confused with profile monitoring, where the observations are profiled modeled as regression models themselves, see Noorossana et al. (2012).

Another difference with the standard setting is that there are no natural subgroups because we are observing a time series. This is called control chart for individuals in the SPC literature, see Montgomery (2005). In this case one cannot use the sample variance of the rational subgroup as it is undefined for groups of size 1. As a proxy, the moving range \( |x_{i+1} - x_i| \) is used as an estimator for the short term standard deviation. This is important in case of a gradual shift in the mean. The limit that is used is a multiple of the moving range.

4.1 Point estimate through linear regression and several methods of determining a confidence interval

One of our goals is to construct a so called overall health index. This would be a single function, letting us know whether something is wrong with the turbine. Such an index might not necessarily provide us with more details on what exactly is wrong regarding the cause of the out-of-control behavior. We will pursue two directions: one based on temperature, one based on power output.

One idea is based on the fact that temperatures are often heavily correlated. If we can make reasonable estimations regarding a temperature variable, based on one or more explanatory variables, then we can indicate when something is getting out of control. We do this by creating sufficiently good regression models, and a way to analyze the deviations from our prediction.

However, not all imminent failures are causing an increment in the temperatures. For this reason, we will construct a second overall health index for the turbines. Informally speaking, the idea is that it is highly unlikely that an imminent failure goes undetected by both these overall health indices. The second index will be based on a prediction model for the power output of the turbine.

We will construct several different temperature models as potential health indices. In general, let

- \( T \), be the temperature variable which we will use as the dependent variable in our regression model, and let
- \( x_1, x_2, \ldots, x_n \), be the independent variables which we will use to predict \( T \). The nature of these variables will be explained in detail later.

We use the built in LM-method in R to construct the best possible prediction for our dependent variable \( T \)

\[
T = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon, \beta_i \in \mathbb{R}. \tag{4.1}
\]

Note that in some cases, for example the nacelle model, \( n = 1 \). This leads to a simplified version \( T = \beta_0 + \beta_1 x_1 + \varepsilon \). Now, for given values of \( x_1, \ldots, x_n \), we can create an estimator of \( T \):

\[
\hat{T} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n. \tag{4.2}
\]
The difference between the estimator and the actual value at that time is denoted as \( \varepsilon \sim \mathcal{N}(0, \sigma^2) \), with unknown \( \sigma^2 \). This is an error term, and it is also known as the residuals of the regression model in equation 4.1. In the case of a single explanatory variable, this model is called simple linear regression.

Thus, we now have \( n \) data points \((x_i, y_i)\), which can be described by a simple linear regression model \( y_i = \alpha + \beta x_i + \varepsilon_i \). The goal is now to find the best fit for a straight line, where best is understood as in the least-squares approach: A line that minimizes the sum of squared residuals of the linear regression model. We can verify that

\[
\hat{Y}_x = \hat{\alpha} + \hat{\beta} x
\]

is the least squares estimator defined by \( \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \) and \( S_{xx} \) the least squares estimator defined by \( \sum_{i=1}^{n} (x_i - \bar{x})^2 \). The formulas above allow us to calculate point estimates of the slope and the intercept of the regression model for a given set of data. However, they do not tell us how precise these estimates are. They may be accurate with a very high level of precision, or quite the opposite. To this end, we use confidence intervals. There are two methods we can use which, as we will show, give us practically the same result.

4.1.1 Confidence interval about the mean response

If we know the mean response for a certain value of \( x \), that is \( \mu_{Y|x} = E(Y|x) = \alpha + \beta x \), we may construct a confidence interval around this value. This is called a confidence interval on the mean response, or a confidence interval about the regression line. We obtain a point estimate of the mean \( \mu_{Y|x_0} \) of \( Y \) at \( x = x_0 \) from the fitted model as \( \hat{\mu}_{Y|x_0} = \hat{\alpha} + \hat{\beta} x_0 \), which is an unbiased estimator of \( \mu_{Y|x_0} \), since \( \hat{\alpha} \) and \( \hat{\beta} \) are unbiased estimators of \( \alpha \) and \( \beta \). We can show that the variance of \( \mu_{Y|x_0} \) is equal to

\[
\text{Var}(\hat{\mu}_{Y|x_0}) = \sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right). \tag{4.5}
\]

Since \( \hat{\mu}_{Y|x_0} \) is normally distributed, we can substitute \( \hat{\sigma}^2 \) for \( \sigma^2 \) to create a random variable which follows a Student-\( t \) distribution with \( n - 2 \) degrees of freedom, as is shown in equation 4.6

\[
\frac{\hat{\mu}_{Y|x_0} - \mu_{Y|x_0}}{\sqrt{\hat{\sigma}^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}}. \tag{4.6}
\]

This leads to the following formula for a confidence interval about the mean response: A 100(1 - \( \alpha \))% confidence interval about the mean response at the value of \( x = x_0 \), say \( \mu_{Y|x_0} \) is given by

\[
\left( \hat{\mu}_{Y|x_0} - t_{\frac{\alpha}{2}, n-2} \sqrt{\frac{\hat{\sigma}^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}}, \hat{\mu}_{Y|x_0} + t_{\frac{\alpha}{2}, n-2} \sqrt{\frac{\hat{\sigma}^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}} \right). \tag{4.7}
\]

Note that this confidence interval does not have the same width across the entire regression line. It gets wider when \( |x_0 - \bar{x}| \) increases, i.e. the uncertainty about the mean response increases when we move away from the mean of the \( x \)-values.

4.1.2 Prediction interval

An important application of a regression model is predicting new or future observations \( y \) corresponding to a specified level of the regressor variable \( x \). If \( x_0 \) is the value of the regressor variable of interest, then \( \hat{y}_0 = \hat{\alpha} + \hat{\beta} x_0 \) is the point estimator of the “new” value of the response. We are interested in obtaining an interval estimate for this future observation \( y_0 \). If we define \( \varepsilon_p = y_0 - \hat{y}_0 \), then \( \varepsilon_p \) is normally distributed with mean zero and variance

\[
\text{Var}(\varepsilon_p) = \sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right). \tag{4.8}
\]
Similar to Section 4.1.1, we can now show that
\[
\frac{\hat{y}_0 - y_0}{\sqrt{\sigma \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}}
\]  
(4.9)
has a Student-t distribution with \(n - 2\) degrees of freedom.

The 100(1 - \(\alpha\))% prediction interval on a future observation \(y_0\) at value \(x_0\) is given by
\[
(\hat{y}_0 - t_{\frac{\alpha}{2}, n-2} \sqrt{\sigma \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}), \quad \hat{y}_0 + t_{\frac{\alpha}{2}, n-2} \sqrt{\sigma \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}.
\]  
(4.10)

### 4.1.3 Shewhart control charts for individual observations

As we explained in Chapter 4, Shewhart charts are statistical control tools, typically used to assure quality control in business or industrial processes. When it is impractical to use individual subgroups, an individual control chart is used. Based on an in-control period, upper and lower control limits of a process are determined. As was explained in Chapter 4, the difference between a data point \(x_i\) and its predecessor \(x_{i-1}\) is calculated as \(MR_i = |x_i - x_{i-1}|\). The mean of these values is calculated as
\[
\overline{MR} = \frac{\sum_{i=2}^{m} MR_i}{m - 1}
\]  
(4.11)
and the control limits at \(x_0\) are defined as
\[
(\hat{y}_0 - k\overline{MR}, \hat{y}_0 + k\overline{MR}),
\]  
(4.12)
where \(k\) depends on the level of confidence used. Since \(\overline{MR}\) is an estimate for the standard deviation, \(k\) is simply the number of sigmas we wish to allow.

### 4.1.4 Comparing Shewhart chart and prediction confidence interval

Though we will not prove this mathematically, we will show visually that the prediction interval and the Shewhart chart are virtually identical. This is mostly due to the fact that our data set consists of so many observations. We will plot both the prediction confidence interval and the Shewhart chart control limits over a scatter plot of the oil temperature versus environmental temperature (Figure 4.2).

In Figure 4.2, the prediction CI’s are plotted in red, and the Shewhart Chart control limits are plotted in blue. Clearly, they coincide almost perfectly, since the blue line overlaps the red one.

### 4.1.5 Level of confidence

The frequency of the warnings is influenced by the imposed level of confidence we use for the warnings. Using a low level of confidence will result in a model that is more sensitive to deviations from the mean. This way, smaller deviations will be picked up, and theoretically any malfunction will be detected faster. However, it also results in a higher number of false warnings. Unfortunately, to the best of our knowledge, nothing concerning a proper level of confidence with this type of predictive maintenance can be found in literature. There is some precedent concerning other types of applications (such as the classical size deviations of machine parts), but there is no reason to assume that the same level of confidence would be applicable here. We choose an average of one false warning per day to be acceptable. There are 360 measurements in a day, so under in-control circumstances, this would mean a confidence level of \(1 - \frac{1}{360}\), or roughly 0.9975.

### 4.2 Identification of in-control period

In order to identify the behavior of various parts as out-of-control, we need to identify an in-control period. This has to be a period without any major mechanical errors. We know a major maintenance event took place on November 16th, so we can already take this date as an upper bound of the in control period. However, out of control behavior most likely started earlier than that date. We would like to identify a period that is as long as possible, without any negative influence of starting out of control behavior prior to the November 16th-maintenance. In order to achieve this, we will look at the correlation coefficient of the environmental temperature and the
oil temperature. We will take the correlation coefficient of these two variables up to a certain point in time. We will then vary this point and check which time corresponds to the highest correlation coefficient.

\[
\rho(t) = \rho_{e,o}(T_0, t), T_0 \leq t \leq T_M. \tag{4.13}
\]

Where \(\rho_{e,o}(T_0, t)\) is defined as the correlation coefficient of the environmental temperature and the oil temperature from time \(T_0\) up to time \(t\). We can do the same thing for the nacelle temperature. They are plotted in Figure 4.3.

The y-axis shows the correlation coefficient. The x-axis shows time. As we expected, the correlation values initially fluctuate, then stabilize, and show a slight decrease leading up to the maintenance event on November 16th. Checking for the maximum shows that this occurs at \(T = 121\), or October 18th. We will take 19/06/2013 up to 12/10/2013 as our in-control period. Interestingly, the correlation coefficient increases after the maintenance.

We can apply the same method to the Nacelle temperature. This is shown in Figure 4.4. This gives us an optimal value at day 2013−10−12. The correlation for the nacelle at 2013−10−12 is 2013−10−12. This is very close, so we take 2013−10−12 to be the limit of our in-control period.

### 4.3 Reducing autocorrelation

Currently, every variable in the data set is measured every four minutes. While this provides us with a great amount of information, it also leads to some issues. In particular the issue of autocorrelation. In order to reduce this, we will look at the data in a 4 hour interval. This should reduce autocorrelation a lot.

We can plot autocorrelation for several different lags at once using the acf-function in R. We will compare the original data set and the reduced data set for bearing temperature and the main generator temperature. This is shown in Figure 4.5.
4.4 Model selection criteria

In order to compare different models with one another, there are several possible statistics we can use. We will take into account the coefficient of determination \((R^2)\), the adjusted \(R^2\), the Akaike’s Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the Mallow’s \(C_p\).

The (traditional) \(R^2\) is defined as

\[
R^2 = 1 - \frac{SS_E}{SS_T}.
\] (4.14)

Which in the case of ordinary least squares regression is equal to

\[
R^2 = \frac{SS_R}{SS_T}.
\] (4.15)

In the above definitions,

\[
SS_T = \sum_i (y_i - \bar{y})^2.
\] (4.16)

\[
SS_R = \sum_i (\hat{y}_i - \bar{y})^2.
\] (4.17)

\[
SS_E = \sum_i (y_i - \hat{y}_i)^2.
\] (4.18)

We see that equation (4.15) holds since in the case of ordinary least squares, we have

\[
SS_T = SS_E + SS_R
\] (4.19)

\(R^2\) takes values between 0 and 1, and can be interpreted as the fraction of variance explained. It is an indicator of how well a regression model fits the data, but it does not take into account the number of parameters in the model. Because the regression models are calculated by minimizing
Figure 4.3: Correlation coefficient of oil temperature and environmental temperature as function of time

$SS_R$. $R^2$ will never decrease when another variable is added to the model. This makes it tempting to keep adding variables when looking at $R^2$ to determine the quality of a model.

The adjusted $R^2$ is an attempt to account for the phenomenon of $R^2$ always increasing when extra variables are added. It is defined as

$$\bar{R}^2 = 1 - \left(1 - R^2\right) \frac{n-1}{n-p-1},$$

where $p$ is the number of explanatory variables in the model (not including the intercept), and $n$ is the sample size. Because of its penalization of the model size, it is a better statistic than $R^2$ to use for model selection.

The AIC is defined as

$$AIC = -2L_m + 2p,$$

The BIC is closely related to the AIC, but is defined as:

$$BIC = -2L_m + p + \log(n).$$

It also takes into account the goodness of fit of the model, as well as the number of parameters. However, the penalty is less severe, since it only adds $p$, as opposed to $2p$ in the AIC.

The Mallow’s $C_p$ is defined as

$$C_p = \frac{SS_E}{S^2} - n + 2p,$$

where $S^2$ is the residual mean after regression of the complete set of $K$ regressors. It can be estimated by the mean square error. $SS_E, n$ and $p$ are defined above.

It can be shown that there is a one-to-one relationship between the AIC and Mallow’s $C_p$ in the case of simple or multiple linear regression. They are equivalent up to a constant. This can be shown by using an equivalent definition of Mallow’s $C_p$. 

32
Figure 4.4: Correlation coefficient of nacelle temperature and environmental temperature as function of time. The green vertical line depicts the in-control moment with the highest correlation coefficient.

We can use matrix notation to denote multiple linear regression, so that the model is defined as follows

\[ Y = X\beta + \sigma \varepsilon. \] (4.24)

where \( Y \) is a random vector in \( \mathbb{R}^n \), \( X \) is the model design matrix containing the observed variables, \( \beta \) is the unknown vector in \( \mathbb{R}^p \) of regression coefficients to be estimated, \( \sigma \) is the noise level and \( \varepsilon \) is a random vector in \( \mathbb{R}^n \) representing model noise. It can be verified that this is equivalent to the definition that was previously used.

Then Mallow’s \( C_p \) can be denoted as:

\[ C_p = \frac{||Y - X\hat{\beta}\|^2}{\hat{\sigma}^2} + 2\hat{df} - n, \] (4.25)

Where \( \hat{\sigma}^2 \) is an estimator of the variance based on the full linear model fitted with the least square estimator \( \hat{\beta}^{LS} \), and \( \hat{df} \) is an estimator of the degrees of freedom of the model, which in our case is equal to \( k \).

Similarly, we can use an alternative definition for the AIC:

\[ \text{AIC} = -2 \sum_{i=1}^{n} \log f(y_i|\hat{\beta}_I) + 2k. \] (4.26)

And finally we define the unbiased estimator of loss

\[ \delta_0(\hat{\beta}) = ||Y - X\hat{\beta}\|^2 + (2\text{div}_Y(X\hat{\beta}) - n)\hat{\sigma}^2. \] (4.27)

So that we have the following link between the AIC and Mallow’s \( C_p \):

\[ \delta_0(\hat{\beta}) = \hat{\sigma}^2 \times C_p(\hat{\beta}) = \hat{\sigma}^2 \times (\text{AIC}(\hat{\beta}) - n), \] (4.28)

which shows that they are equivalent up to a constant. This relationship is further explained in Boisbunon et al. (2014).
4.5 Conclusions

- Several historical methods exist that we can make use of. None of the methods exactly describe the situation in this report, but combining them, we can create a method that is applicable to our situation. In particular, we make use of linear regression to create estimates of temperature and power variables and then make use of Shewhart’s methods for process control to calculate control limits. In case of simple linear regression, it is also feasible to use the prediction interval instead of Shewhart control limits. The differences between the two are negligible.

- We can identify the in-control period by looking at the correlation between the oil temperature. While the turbine is behaving in-control, this correlation is higher than when the turbine exhibits out-of-control behavior. Plotting the correlation as a function of time and identifying the peak gives us a good estimate of what moment to use for our in-control limit.

- Autocorrelation may cause a problem for the accuracy of the temperature models. We can reduce the autocorrelation by looking at a reduced data set which consists of measurements once every 4 hours instead of every 4 minutes.

- Several performance measures exist to check the quality of regression models. Because we want to include a penalty for the amount of variables in the model (to avoid overfitting) we use the adjusted $R^2$ and the BIC. In some cases it is easier to implement using Mallow’s $C_p$. This is equivalent, since there exists a one-to-one relation between the two.
Chapter 5

Models

5.1 Temperature models

We now apply the above methods to several temperature variables, namely:

- The nacelle temperature
- The oil temperature
- The bearing temperature
- The gearbox temperature
- The main generator temperature
- The starting generator temperature.

For each of these variables, we create a regression model and construct a method to check for warnings. We do this using the methods described in Chapter 4. Concretely, for each variable we do the following:

First, we determine an in-control period. In most cases, we will use the period derived in section 4.3, which is 19/06/2013 up to 12/10/2013. Using this in-control period, we construct two preliminary regression models. The first uses all data in the in-control period, the second uses a subset of that data which is selected once every 4 hours instead of once every 4 minutes. Using several criteria, we determine which of the two regression models is most suitable to continue our analysis with. To this end, we will use a visual inspection of an autocorrelation plot of the residuals, as well as the $R^2$ and BIC values of the models. We take the model that we selected in the previous step, and further refine it by adding and removing variables. This is done either through an all-subset selection method or through a stepwise selection method, using the BIC as a selection criterion. When we have the best possible selection method, we calculate the predicted values and residuals for the entire data period. For these residuals, we create a Shewhart control chart and determine which observations are beyond the control limits. These are then classified as warnings and an accompanying plot which depicts which temperature values are out of control is created. The results are then summarized and conclusions are drawn from each model.

5.1.1 Regression model for the nacelle temperature

We first try our approach on the nacelle temperature. Since the nacelle contains all the other components, its temperature should reflect a change in behavior in any of them. It has the added benefit of being the easiest to model, since, as we’ve seen, it follows the environmental temperature quite nicely. In fact, the environmental temperature is the only explanatory variable in the regression model. The first thing we need to do is determine the in-control period. We have already determined a global in-control limit in section 4.3, which is set at 16/10/2013. This is indeed the limit we use for our in-control period in this case as well.

Next, we compare whether a regression model based on data every 4 hours or every 4 minutes would give us the best results. To this end, we create two preliminary regression models. One based on an in-control period and using all data points within that period, and one using a subset of the data within that period, with measurements once every 4 hours instead of every 4 minutes. Earlier, see Figure 4.5, we noticed that autocorrelation in the raw temperature data
drops significantly if we look at data in larger time intervals. In the cases of both generator temperature and bearing temperature, we see a steady decline in approximately the first 36 hours, after which the correlation coefficient somewhat stabilizes. We would now like to see if the same holds true for the residuals of the nacelle temperature. To this end, we create an autocorrelation plot, depicted in Figure 5.1, and visually inspect it to see which model performs best in terms of autocorrelation. Unsurprisingly, the autocorrelation in the residuals of the measurements and predictions taken every 4-hours is much smaller. Within the in-control period, there is a certain period with a lot of missing values. In this case, we want to see the difference in autocorrelation between the 4-minute and 4-hour intervals. Since this difference is usually quite large, it is okay to simply use the data up to where the missing values start to occur and use that to create the plot. This is indeed what we have done in Figure 5.1.

Furthermore, we can compare the $R^2$ and BIC values for the two models as well. These are listed in Table 5.1. We see that the $R^2$ values of the two models are almost identical, but the BIC value is much smaller for the 4-hour model. Based on this result, it is best to create a regression model based on the 4-hour data.

Since there is only one explanatory variable in the model, in this case it is not necessary to use stepwise- or all subset regression to improve the selection of the explanatory variables. We can simply use the regression model as it was given by the linear model function in R. It would be possible for a selection algorithm to create a model either without the intercept coefficient or without the regression variable, but observing the $p$-values for both variables, we can conclude that this will not be the case, so we do not need to test for it.

Table 5.1: Comparison of $R^2$ and BIC values for the nacelle temperature regression models using 4 minute or 4 hour intervals

<table>
<thead>
<tr>
<th></th>
<th>4 minute interval</th>
<th>4 hour interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.918511</td>
<td>0.8880073</td>
</tr>
<tr>
<td>BIC</td>
<td>2, 888.4213876</td>
<td>2, 58.3085082</td>
</tr>
</tbody>
</table>
The coefficients of this model are listed in Table 5.2.

|                     | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------------|----------|------------|---------|---------|
| (Intercept)         | 9.0321   | 0.6184     | 14.61   | 0.0000  |
| dataOper3InControl | 0.8849   | 0.0280     | 31.61   | 0.0000  |
| EnvTemp             |          |            |         |         |

Table 5.2: Regression model of the nacelle temperature, 4 hour intervals

Using this regression model, which is based on the in-control period, we can predict point estimates for the nacelle temperature for the entire data period. These residuals are then used to calculate the control limits for a Shewhart chart, which allows us to see which data points behave out-of-control. The control limits were calculated using data within the in-control period only. Any data acquired after the control limit did not play a part in the computation of the control limits. The control limits of the in-control period, however, are still used to determine whether a data point after the control limit is out-of-control. The Shewhart chart is shown in Figure 5.2. Although due to the implementation of the Shewhart chart in R it is not possible to see the exact dates and/or times of the data point, we can see that there are relatively few out-of-control data point at the beginning, while after some time, more points behave out-of-control. In total, 1914 measurements are beyond the control limits, out of a total of 30775, or approximately 6.2%. This is much more than is to be expected from a well-behaved system using 3-sigma control limits, and this is definitely a warning that something is not functioning properly within the turbine.

We can now create a scatter plot of the nacelle temperature versus the environmental temperature. In this scatter plot, we can depict which measurements are out-of-control and which are in-control. This shows us that nearly all of the out-of-control points are beyond the upper control limit rather than the lower control limit. Since a defect is more likely to cause an increase in temperature than a decrease, this might indeed be the effect of a malfunctioning of some part of the turbine. We also see that most out-of-control measurements occur at moderate...
temperatures rather than at temperatures which are extremely high or extremely low. The lack of out-of-control measurements at lower temperatures may however also be caused by the period that we identified as in-control. This period, which runs from June to October, did not have many freezing temperatures, so the predictions may be less accurate in months during which the environmental temperature is low. An in-control period that covers both warm and cold months might yield even better results, but unfortunately this is not possible with the data that we currently have. The scatter plot is shown in Figure 5.3.

If we map the out-of-control measurements to a plot of the nacelle temperature versus time, as is shown in 5.4, we can indeed see that the time after which more out-of-control measurements start to show up coincides quite well with the large maintenance of the turbine on November 16, 2013. This confirms our initial suspicion that this is where the malfunctioning of the turbine initially started.
5.1.2 Regression model for the oil temperature

We now create a model like the previous one for the oil temperature. This is done in a similar way to the nacelle model. Since we have seen a very strong correlation between the environmental temperature and the oil temperature, just as we did for the nacelle temperature, we once again construct a model that uses the environmental temperature as its only explanatory variable. Furthermore, we once again use the same in-control period, which is cut off at 16/10/2013.

Our first step is to determine whether we will use the 4-hour or the 4-minute period. To this end, we create and visually compare the autocorrelation plots for both the 4-hour and the 4-minute data set. We see that the autocorrelation in case of the 4-hour data set is much smaller. The autocorrelation plot is depicted in Figure 5.5.

Apart from the autocorrelation, we also look at the $R^2$ and the BIC values. We see that the $R^2$ is very slightly lower for the 4-hour model. It has a value of 0.592 for the 4-minute model, and a value of 0.579 for the 4-hour model. The BIC, however, is much lower for the 4-hour model. It has a value of 1589 for the 4-minute model, and a value of 264 for the 4-hour model. The $R$-squared and BIC-statistics for the model are listed in Table 5.3.

Based on these values as well as the autocorrelation plot. We determine that it is better to use the 4-hour data set for the creation of our linear model. Since we are once again using simple linear regression, and the $p$-values of both coefficients in the initial regression model are sufficiently low, we do not need to perform a selection algorithm to determine the final regression model. The coefficients of the regression model for the oil temperature are printed in Table 5.4.

<table>
<thead>
<tr>
<th></th>
<th>4 minute interval</th>
<th>4 hour interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.5924421</td>
<td>0.5798339</td>
</tr>
<tr>
<td>BIC</td>
<td>2, 1.580928 × 10^4</td>
<td>2, 264.9544841</td>
</tr>
</tbody>
</table>

Table 5.3: Comparison of $R^2$ and BIC values for the oil temperature
Using this regression model, we predict point estimates for the oil temperature for each available data point. We then compute the residuals of the regression model, and use these to compute the Shewhart control limits. The control limits are again computed using data from the in-control period only. This allows us to see which measurements show an out-of-control behavior. The chart, which looks quite similar to the Shewhart chart produced for the nacelle temperature (Figure 5.2), is mostly in-control for the first part, and then starts to show more out-of-control observations.

Again, most of the out-of-control measurements are beyond the upper control limit, rather than beyond the lower control limit. We can easily see this if we create a scatter plot of the environmental temperature and the oil temperature, in which we add lines to depict the control limits, and color-characterize the data points which are out-of-control. This plot is depicted in Figure 5.7. In total, the number of points beyond the control limits is 1407 out of a total of 30775 measurements, or approximately 4.5%. Once again, this is much more than is to be expected from a Shewhart chart which uses a 3 sigma level of confidence, and is definitely a sign that something is not functioning properly.

We can once again create a plot of the oil temperature versus time, in which we color-characterize the out-of-control data points. This shows us that the oil temperature starts to behave out-of-control right after the start of the in-control period and the amount of out-of-control measurements does not seem to decrease.
Figure 5.6: Shewhart chart for the residuals of the oil temperature, using 4 minute intervals
Figure 5.7: Scatter plot of oil temperature and environmental temperature, with lines representing the point estimates (green) and the prediction interval (red). The points which are classified as warnings according to the prediction interval are color-characterized in red.
Figure 5.8: Oil temperature versus time. The points which are classified as warnings according to the Shewhart chart are color-characterized in red.
5.1.3 Regression model for the bearing temperature

Next, we apply our method to the bearing temperature. This is done in a slightly different way, because the bearing temperature cannot be explained by a single regression variable. For this reason, we construct a linear model using multiple regression. After visual inspection, we note that the following variables are the most likely candidates to be good explanatory variables:

- Wind Speed
- Generator speed
- Environmental temperature

It must be noted that the power output is also a potential explanatory variable. However, due to the number of missing values, it would limit the in-control period too much to be able to create a proper estimate of in-control behavior. For that reason, the power output is omitted from this model, as well as the gearbox and main generator model. Other temperature variables, such as the generator temperature, or the oil temperature, cannot be used. The reason for this is that the bearing temperature most likely influences other part temperatures, because of heat transfer through the air, to parts in its proximity. This means that, if the bearing temperature would rise because of some mechanical error, this heat transfer would then cause the other part temperatures to increase as well. However, since in this case the bearing would be the source of the heat, it would not be correct to classify these other temperatures as explanatory variables. This correlation effect is seen most strongly with the gearbox temperature, as we have already shown in Figure 3.2 in chapter 3. For now, we also leave out any vibration variables. We will later include these in the model for the gearbox temperature. We do not need to add the vibration variables in both cases, since the bearing and the gearbox temperature are so similar. This allows us to see what impact these variables have on the model. All this leaves us with 3 independent, explanatory variables. For each variable, up to the third power will be used. This, along with the intercept variable, leaves us with 10 regression variables and correspondingly 10 linear coefficients $\beta_0, \beta_1, ..., \beta_9$. Our in-control limit is defined identically to the way it was defined before, which is 16/10/2013. With these 13 coefficients, we construct an initial regression model that is filled completely with all 10 of them. We then use the function “leaps” in R to perform an all subset selection, which uses Mallow’s $C_p$ as a selection criterion. As we have shown in chapter 4, Mallow’s $C_p$ shares a one-to-one relation with the AIC, and is also closely related to the BIC. For all possible subsets of the 10 explanatory variables, the leaps function calculates the best possible regression model, using the least sum of squares, and compares these models using Mallow’s $C_p$. Note that in the case of 10 variables, this amounts to $2^{10}$ or 1024 regression models, from which one is selected. The function with the lowest Mallow’s $C_p$ is then used as the regression model to perform the rest of our calculations with.

The above (computing the best possible regression model) is done for both the 4-minute data set and the 4-hour data set. This gives us two different regression models, for which we will visually compare the autocorrelation plots, and inspect the $R^2$ and BIC values. The autocorrelation plot, which is shown in Figure 5.9, once again shows a great decrease in values for the 4-hour model.

The $R^2$ and BIC values for the bearing temperature are listed in Table 5.5. We see that the $R^2$ value is much greater for the 4-hour model, and the BIC is much smaller for the four-hour model.

<table>
<thead>
<tr>
<th></th>
<th>4 minute interval</th>
<th>4 hour interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.5319582</td>
<td>0.8110542</td>
</tr>
<tr>
<td>BIC</td>
<td>10, 5970.8993848</td>
<td>7, 1392.8522837</td>
</tr>
</tbody>
</table>

Table 5.5: Comparison of $R^2$ and BIC values for the bearing model, using 4 minute or 4 hour intervals
Figure 5.9: Autocorrelation for the bearing temperature, using the entire data set in the top picture and the reduced data set (4-hour intervals) in the bottom picture.

Table 5.6: Regression model for the bearing temperature, four hour intervals

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 36.4283 | 0.8228 | 44.27 | 0.0000 |
| I(EnvTemp) | 0.2244 | 0.0284 | 7.90 | 0.0000 |
| I(WindSpeed) | 0.6687 | 0.4418 | 1.51 | 0.1306 |
| I(WindSpeed^2) | 0.2122 | 0.0658 | 3.22 | 0.0013 |
| I(WindSpeed^3) | -0.0131 | 0.0029 | -4.47 | 0.0000 |
| I(GenSpeed^2) | 0.0000 | 0.0000 | 8.92 | 0.0000 |
| I(GenSpeed^3) | -0.0000 | 0.0000 | -5.99 | 0.0000 |

in Figure 5.10, we see a great increase of out-of-control points after a certain amount of time. Again, most out-of-control limits are beyond the upper control limit rather than the lower limit.

The number of out-of-control measurements is now even greater than it was before. There are 536 measurements beyond the control limits, out of a total of 30775 measurements. This means that approximately 1.7% of the total measurements are beyond the control limits. This is more than is to be expected, and a definite sign that something is wrong.

Because this is a multiple regression model, it is not meaningful to create a scatter plot for the bearing temperature. Since a scatter plot purely aids in visual inspection of the results, and is in no way vital to our model, we can safely skip it. We will now construct a plot of the bearing temperature versus time. This plot is depicted in Figure 5.11. We see that initially, during the in-control period, the bearing temperature only spans a relatively narrow range of temperatures. Almost all of the measurements are roughly within the 60 to 70-degree span. When we move beyond the control limit, we see that this band gradually increases in both directions. By the time the first major maintenance event took place, which is depicted by the first red line in the plot, the temperatures span across a band at least twice as wide. This behavior only worsens as time increases. With this behavior also come more out-of-control points, and we indeed see that
virtually all out-of-control measurements are beyond the large maintenance event.

The sudden increase of out-of-control measurements coincides with the time the nacelle temperature and oil temperature started to behave out-of-control. This further strengthens our suspicions that around the large maintenance event at November 16, 2013, the turbine first showed signs of malfunctioning.
Next, we apply our methodology to the gearbox temperature. Just as with the bearing temperature, the gearbox temperature cannot be explained by a single regression variable, which means that once again we will use multiple regression to construct a good regression model. Apart from all the variables we used for the bearing temperature regression model, we will this time also use some vibration readings which are sufficiently well behaved. These are GbxMSA10KHzOvr, GbxMSA1KHzOvr, GbxMSD500HzOvr, GbxOSA10KHzOvr, GbxOSA2KHzOvr, and GbxOSD1KHzOvr. This gives us a total of 18 explanatory variables and, including the intercept, 19 coefficients for the regression model to be determined. Note that other temperature variables are still excluded, since an increase in a part temperature may be caused by an increase in the gearbox temperature, and although this would imply a correlation, it would be incorrect to classify these variables as explanatory variables.

We first construct two initial regression models using all 19 variables, using the 4-minute data set in for the first model and the 4-hour data set for the second. We then apply an all-subset selection method, using Mallow’s $C_p$ as our selection criterion. This gives us two regression models, which we can once again compare by visual inspection of the autocorrelation plot, and by looking at the BIC- and $R^2$ values. We see that the autocorrelation once again decreases massively for the 4-hour residuals. The autocorrelation plot is depicted in Figure 5.12.

When looking at the $R^2$ and BIC values, which are listed in Table 5.7, we see that the $R^2$ value increases substantially when looking at the 4-hour model. The BIC value is also substantially smaller, dropping from 2038 to only 15. We also see that the 4-hour model only uses 5 explanatory variables, whereas the 4-minute model uses 13. This means that by all our criteria, the 4-hour model is the superior one. It is also the simpler model of the two, which is also desirable.

In the 4-hour model, we see that only 4 variables and an intercept are included in the model. This result is quite surprising, as most of the variables which were in the initial bearing model, are also in the initial gearbox model. It seems the vibration variables that were added explain the behavior so well, that a lot of other variables are excluded. The coefficients and variables
Figure 5.12: Autocorrelation for the gearbox temperature, using the entire data set in the top picture and the reduced data set (4-hour intervals) in the bottom picture.

Table 5.7: Comparison of $R^2$ and BIC values for the gearbox temperature, using 4 minute and 4 hour intervals

<table>
<thead>
<tr>
<th></th>
<th>4 minute interval</th>
<th>4 hour interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.640309</td>
<td>0.736719</td>
</tr>
<tr>
<td>BIC</td>
<td>13, 2150.7955914</td>
<td>9, 996.2786182</td>
</tr>
</tbody>
</table>

which are included in the final model are listed in Table 5.8.

|                      | Estimate  | Std. Error | t value | Pr(>|t|) |
|----------------------|-----------|------------|---------|---------|
| (Intercept)          | 40.4468   | 0.8258     | 48.98   | 0.0000  |
| I(EnvTemp^3)        | 0.0001    | 0.0000     | 5.23    | 0.0000  |
| I(WindSpeed^2)      | 0.0760    | 0.0115     | 6.62    | 0.0000  |
| I(GenSpeed)         | 0.0138    | 0.0025     | 5.49    | 0.0000  |
| I(GenSpeed^2)       | -0.0000   | 0.0000     | -2.25   | 0.0247  |
| GbxMSD500HzOvr      | 47.8100   | 15.8715    | 3.01    | 0.0027  |
| GbxOSA10KHzOvr      | 44.3246   | 6.2127     | 7.13    | 0.0000  |
| GbxOSA2KHzOvr       | 2.8273    | 0.7810     | 3.62    | 0.0003  |
| GbxOSD1KHzOvr       | -221.9257 | 26.0790    | -8.51   | 0.0000  |

Table 5.8: Regression model for the gearbox temperature, 4 hour intervals

Just like with the bearing model, the residuals are then plotted, and a warning level is determined. We can then color-code the warning points in a graph of gearbox temperature versus time.

We once again use this regression model to determine the residuals of the gearbox temperature, and construct the control limits. A Shewhart chart of the residuals is plotted in Figure 5.10.

We see that the residuals are much more well-behaved than those of the bearing temperature,
even though the temperatures behave very similar. In case of the gearbox temperature, 21382 out of a total of 30775 measurements are beyond the control limits, or 4.8%. This is again beyond the number that should fall outside the 99.75% control limits in an in-control situation. If we plot the gearbox temperature versus time, we see similar behavior as we saw with the bearing temperature. Again, the temperatures start off in a relatively narrow band, and this gradually widens. As is to be expected, we start to see out-of-control behavior when the wider range of temperatures occur. This again seems to coincide with the maintenance on November 16, 2013.
5.1.5 Regression model for the main generator temperature

For the main generator, we perform the same steps as we did for the bearing and gearbox temperature. We once again use multiple linear regression to construct a regression model, since there is no single independent explanatory variable which we can use to explain the temperature behavior. For the initial model, we use the variables:

- Environmental Temperature
- Wind Speed
- Generator Speed
- Gearbox Temperature

For both the 4-hour and 4-minute data set, we create the initial regression models, which include all explanatory variables. We then perform an all-subset selection method, which uses Mallow’s $C_p$ as its criterion, and compare the autocorrelation plots and the BIC and $R^2$ values. The autocorrelation plot shows us much lower values for the 4-hour model. This is consistent with earlier observations in other models. The autocorrelation plot is depicted in Figure 5.15.

The $R^2$ and BIC values for both models are listed in Table 5.9. We see that the $R^2$ drops quite significantly, from 0.79 to 0.59. Both models use 5 variables, and for the 4-hour model, the BIC is significantly lower. This means that despite the higher $R^2$ value of the 4-minute model, we will use the 4-hour model, as it performs better in terms of autocorrelation and BIC-value.

The coefficients and variables for the main generator regression model are listed in Table 5.10. We see that the environmental temperature, the wind speed, the generator speed and the power are included. All have $p$-values which are quite high, especially the environmental temperature, but in terms of Mallow’s $C_p$, it is still an improvement that they are included.

We then use this model to predict point estimates for all points in the data set, and with them compute the residuals. The residuals that correspond to the in-control period are used to
Figure 5.15: Autocorrelation plot for the main generator temperature, using the entire data set in the top plot, and the reduced (4h-intervals) in the bottom plot.

<table>
<thead>
<tr>
<th></th>
<th>4 minute interval</th>
<th>4 hour interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.6358048</td>
<td>0.5932789</td>
</tr>
<tr>
<td>BIC</td>
<td>$4 \times 6.7810498 \times 10^5$</td>
<td>4, 252.4814437</td>
</tr>
</tbody>
</table>

Table 5.9: Comparison of $R^2$ and BIC values for the main generator temperature, using 4 minute and 4 hour intervals.

The above data shows that the autocorrelation is stronger in the 4 minute interval than in the 4 hour interval. This indicates a higher degree of dependency in the short term, which is expected given the nature of the data.

Since the main generator is the part of the turbine that we know for a fact malfunctioned, the fact that out-of-control behavior occurs is not surprising. In this case we see data points beyond both the upper control limit and beyond the lower control limit. 788 measurements are beyond the control limit, and in total there are 30775 data points. This means that approximately 2.5% of the measurements is out-of-control.

When plotting the main generator temperature versus time and color characterizing the data points which are beyond the Shewhart control limits, we see that there are some out-of-control data points within the in-control period, but there are a lot more in the out-of-control period. This seems intuitively correct, as there are a lot of temperature spikes in that period, some even outside the control limits.

|                | Estimate   | Std. Error | t value | Pr(>|t|) |
|----------------|------------|------------|---------|----------|
| (Intercept)    | -1352.6316 | 713.6100   | -1.90   | 0.0642   |
| EnvTemp        | -0.0622    | 0.5175     | -0.12   | 0.9049   |
| WindSpeed      | 3.1913     | 1.7221     | 1.85    | 0.0701   |
| GenSpeed       | 0.9154     | 0.4772     | 1.92    | 0.0612   |

Table 5.10: Regression model for the temperature of the main generator.
up to around 130 degrees. These spikes also seem to be where most out-of-control behavior occurs. We can interpret this as a clear warning, and this further reinforces our beliefs that the first malfunctioning occurred around the time of the major turbine maintenance on November 16, 2013.

We can repeat this method for the reduced data set which only measures every 4 hours:

5.1.6  Regression model for the starting generator temperature

The final part for which we create a temperature-based model is the starting generator. Its temperature also cannot be well explained by a single explanatory variable, so we once again construct a multiple regression model in order to explain its behavior. For the initial model, we use the following variables:

- Environmental temperature
- Wind Speed
- Generator Speed
- Power

Again, we create regression models based on both the 4-hour and 4-minute data set. We start with an initial regression model containing all four explanatory variables, and perform all-subset regression, using Mallow’s $C_p$ as a criterion. We compare the autocorrelation plots, and compute the $R^2$ and BIC values. The autocorrelation plot shows us that the autocorrelation is again much smaller for the 4-hour model. The autocorrelation plot is shown in Figure 5.18.

Surprisingly, we see a very large drop in the $R^2$ value of the 4-hour model. Where the 4-minute model has an $R^2$ value of 0.479, the 4-hour model only has an $R^2$ value of 0.011. This is an extremely low value, and practically renders the model unusable. Upon closer inspection, it seems that our selection criteria for the in-control data set left us with only 67 measurements. This may
Figure 5.17: Main generator temperature. Warnings are color-characterized in red.

<table>
<thead>
<tr>
<th></th>
<th>4 minute interval</th>
<th>4 hour interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.479658</td>
<td>0.0112332</td>
</tr>
<tr>
<td>BIC</td>
<td>5.8530654 x 10^4</td>
<td>3.3980901 x 10^4</td>
</tr>
</tbody>
</table>

Table 5.11: Comparison of $R^2$ and BIC values for the starting generator temperature, using 4 minute and 4 hour intervals

be the reason for the poor performance of the model. The BIC did not improve much for the 4-hour model either, which is the reason that despite the higher autocorrelation, we will continue using the 4-minute model in this case.

The coefficients of the 4-minute model after all-subset selection are shown in Figure 5.12.

|                  | Estimate | Std. Error | t value | Pr(>|t|) |
|------------------|----------|------------|---------|---------|
| (Intercept)      | -286.3987| 35.4663    | -8.08   | 0.0000  |
| EnvTemp          | 0.7528   | 0.0476     | 15.82   | 0.0000  |
| WindSpeed        | 2.2242   | 0.3661     | 6.07    | 0.0000  |
| GenSpeed         | 0.2623   | 0.0309     | 8.50    | 0.0000  |
| Power            | 0.1137   | 0.0099     | 11.50   | 0.0000  |

Table 5.12: Regression model for the temperature of the starting generator, four minute intervals

We see that the environmental temperature, wind speed, generator speed and power are all included in the model. We now use this regression model to calculate the residuals of the starting generator temperature, which are then used to calculate the control limits. A Shewhart chart of the residuals is plotted in Figure 5.19. We see that the values are relatively well behaved within the in-control period. In fact, only 7 values in the in-control period are beyond the control limits. As soon as the in-control limit ends, however, there are a lot of out-of-control values. Beyond the point where the main generator was disabled, there is a massive amount of out-of-control values, most of which are beyond the lower control limit. This makes sense,
however, because the regression model is partly dependent on wind speed. During the time where the main generator was active, which includes the in-control period, the starting generator was disconnected at wind speeds of 7 or higher. When the main generator was disconnected however, the starting generator was also used at higher wind speeds. Because the wind speed is included in the regression model, the higher wind speeds lead to unrealistically high predictions, and therefore low residuals. The usage and behavior of the starting generator is so different when it is also used as the main generator, that this regression model cannot predict values properly in this case.

We can still plot the values of the starting generator against time and color code the out-of-control values. This is done in Figure 5.20. We see that the values are very well behaved, apart from the periods where the main generator was disabled. This leads us to believe that there is no issue with the starting generator.
Figure 5.19: Shewhart chart for the residuals of the starting generator temperature
Figure 5.20: Starting generator temperature. Warnings are color-characterized in red.
5.2 Power models

Apart from the health indexes based on the temperatures of several parts, it may also be possible to create an overall health index based on the power output of the turbine. If we can create an estimate for what the power output at a given time should be, then we can analyze the difference between that estimate and the actual power output. Whenever this difference is very large, this could be interpreted as a warning. Fortunately, theoretical models exist for the power output in wind turbines. These theoretical models are based on the wind speed. The theoretical curve for the Vestas V47 turbine is shown in Figure 5.21.

![Theoretical power curve for the Vestas V47](image)

Notice that the cubic shape of this curve resembles the scatter plot of the power output and wind speed in Figure 3.9 in section 3.1. In particular, note three important things about the curve.

- Below 3.5 m/s the turbine produces no power. There is insufficient torque exerted by the wind on the rotor to make it rotate and generate power. We say that 3.5 m/s is the cut-in speed.
- Between 3.5 and 14 m/s, power behaves as a cubic function of wind speed. At 14 m/s, the limit of the generator power output is reached. After this point, the power output no longer increases. We say that 14 m/s is the rated output speed.
- At wind speeds over 25 m/s, the forces exerted on the rotor are so great that there is a serious risk of structural damage. Therefore, the breaking system brings the system to a halt. We say that 25 m/s is the cut-out speed.

Between the cut-in speed and the rated output speed, the estimated output power can be shown to be estimated by the following formula:

\[
W = \frac{1}{2} \rho U^3 \pi d^2 \frac{d}{4}
\]

where \(W\) is the power in watts, \(U\) is the wind speed in meters per second, \(\rho\) is the aerial density in kilograms per cubic meter, and \(d\) the rotor diameter in meters. The power output behaves very differently for the main generator and the starting generator. For this reason, we choose to divide the data based upon which generator is in use, and create a separate regression model for the power output for each generator. Furthermore, we limit the data to entries with a wind speed between the cut-in speed and the cut-out speed.

5.2.1 Power output for the main generator

Since the power output is well known to be a cubic function of the wind speed, we only use the wind speed as explanatory variable. First, we check if the four hour data set or the four-minute
Figure 5.22: Autocorrelation for the power output of the main generator, using the entire data set in the top picture and the reduced data set (4-hour intervals) in the bottom picture.

<table>
<thead>
<tr>
<th></th>
<th>4 minute interval</th>
<th>4 hour interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.8603447</td>
<td>0.9035182</td>
</tr>
<tr>
<td>BIC</td>
<td>$1.4341766 \times 10^4$</td>
<td>3,552.9673935</td>
</tr>
</tbody>
</table>

Table 5.13: Comparison of $R^2$ and BIC values for the power output of the main generator, using 4 minute and 4 hour intervals

data set provides us with a better model. To this end, we create regression models for both cases. We use all-subset regression to create the best possible regression model, using Mallow’s $C_p$ as a performance measure. We then check the autocorrelation on the residuals for both. The autocorrelation plot is shown in Figure 5.22. In this case, the autocorrelation is much less than it was for any of the temperature residuals. Even after four minutes, the autocorrelation immediately drops to below 0.1. In cases where the auto-correlation is not an issue, it is preferable to use the entire data set, since otherwise we discard information without good reason. Apart from the autocorrelation, we also check the $R^2$ and the BIC-value. These values are shown in Table 5.13.

The adjusted $R^2$ increases slightly when we use the 4-hour data set to create the regression model. Furthermore, Mallow’s $C_p$ is a lot smaller. For this reason, we choose to use the 4-hour model in further calculations. The regression coefficients are listed in Table 5.14 and the adjusted $R^2$ and Mallow’s $C_p$ are listed in Table 5.13.

Using this regression model, we predict point estimates for each data point. We then compute the residuals of the regression model, and use the residuals to compute the Shewhart control limits. The control limits are measured using the in-control period as defined in Section 4. The Shewhart chart allows us to monitor which measurements show an out-of-control behavior. The Shewhart chart is plotted in Figure 5.23.

We see that the out-of-control behavior only starts after the in-control period ends. This reinforces our hypothesis that the defect of the main generator occurred around that time. A
|                         | Estimate | Std. Error | t value | Pr(>|t|) |
|-------------------------|----------|------------|---------|----------|
| (Intercept)             | -62.3927 | 31.1523    | -2.00   | 0.0487   |
| I(WindSpeed)            | 16.3277  | 7.0878     | 2.30    | 0.0239   |
| I(WindSpeed^3)          | 0.2504   | 0.0471     | 5.32    | 0.0000   |

Table 5.14: Regression model of the power output, using data from the main generator and 4 minute intervals

Figure 5.23: Shewhart chart for the residuals of the power output, using data for the main generator only

plot of the power output of the main generator versus time is shown in Figure 5.20.

5.2.2 Power output for the starting generator

For the starting generator, we can repeat this process. We once again start by creating two models. One for the four-hour data set, and one for the four-minute data set. On each of these, we use all-subset regression to select the best possible model. Mallow’s $C_p$ is used as a performance measure. We then start out by comparing the autocorrelation of the residuals for each of these models.

Similar to what we saw for the power output of the main generator, the autocorrelation is much less than it was for any of the temperature residuals. Even after four minutes, the autocorrelation immediately drops to below 0.1. Therefore, it is again preferable to use the entire data set. We also check the $R^2$ and the BIC-value. These values are shown in Table 5.15.

We see a slight rise in the adjusted $R^2$, but a significant drop in the BIC. For this reason, we elect to use the four-hour model over the four-minute model. The coefficients of this model are as presented in Table 5.16.

Using this regression model, we predict point estimates for each data point. We then compute the residuals of the regression model, and use the residuals to compute the Shewhart control limits. The control limits are measured using the in-control period as defined in Section 4. The
Figure 5.24: Power output for the main generator. Warnings are color-characterized in red.

<table>
<thead>
<tr>
<th>4 minute interval</th>
<th>4 hour interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.8985457</td>
</tr>
<tr>
<td>BIC</td>
<td>1.0421072 × 10^4</td>
</tr>
</tbody>
</table>

Table 5.15: Comparison of $R^2$ and BIC values for the power output of the main generator, using 4 minute and 4 hour intervals

Shewhart chart allows us to monitor which measurements show an out-of-control behavior. The Shewhart chart is plotted in Figure 5.23, and the plot of the power output versus time is shown in Figure 5.27. In this case, the out-of-control measurements are distributed much more evenly across the entire period. This means that from this model, we are not able to identify any defects. It should be denoted however, that due to the prevalence of NA’s in the measurements of the power output of the starting generator, there is a relatively limited amount of data to work with. It would be advisable to re-identify a (preferably longer) in-control period for the starting generator, and use that period to re-construct the model, as this would lead to more reliable results

|                          | Estimate | Std. Error | t value | Pr(>|t|) |
|--------------------------|----------|------------|---------|---------|
| (Intercept)              | -134.4357| 12.0897    | -11.12  | 0.0000  |
| I(WindSpeed)             | 33.7904  | 2.6203     | 12.90   | 0.0000  |
| I(WindSpeed^3)           | 0.1454   | 0.0136     | 10.72   | 0.0000  |

Table 5.16: Regression model of the power output, using data from the starting generator and 4 minute intervals
Figure 5.25: Autocorrelation for the power output of the starting generator temperature, using the entire data set in the top picture and the reduced data set (4-hour intervals) in the bottom picture.

5.3 Conclusions

All models, with the exception of the model based on the power output of the starting generator, show a clear change in behavior after the end of the in-control period that we identified in Chapter 4. The models show relatively few out-of-control measurements within the in-control period, and subsequently exhibit out-of-control behavior for the remainder of the data period. The fact that all of these models exhibit the same kind of behavior, which is in accordance to what we can expect from the maintenance logs, is a clear indication that our methods accurately depict the health of the wind turbine as a whole. As of now, it is unclear whether we are able to use similar methods to narrow the issue down to a single part of the turbine. All of the temperature models show a similar image, and do not point to a single part as the cause of the problem. The models based on the power output are also interesting. The power output of the main generator shows an image similar to the temperature models, which is to be expected, given the fact that the main generator is the one that failed. However, the power output of the starting generator shows no change of behavior. This may indicate that the power output of the starting generator is simply not affected by the main generator. In that case, the power output may very well still be a viable health index. It would simply mean that the starting generator is still in good health, and no abnormal amount of warnings was shown because of it.
Figure 5.26: Shewhart chart for the residuals of the power output, using data for the starting generator only.
Figure 5.27: Power output for the main generator. Warnings are color-characterized in red.
Chapter 6

Results and conclusions

6.1 Summary

In this report, we were able to use a statistical approach to wind turbine condition monitoring to create a system which monitors the overall health of the turbine. We identified critical relations between variables of different natures, and used these relations to construct models which are able to identify out-of-control behavior. In order to construct these models, we made use of several existing statistical methods and combined them to be suited for our applications. In particular, we applied Shewhart's methods for identifying out-of-control behavior to the residuals of a linear regression model. The models seem to reflect the condition of the wind turbine reasonably well. All the models start to exhibit warnings around October 2013 which is, as we know from the maintenance logs, the time that the main generator first failed. The exception to this rule is the model based on the power output of the starting generator, but this may be because it reflect the health of the starting generator, rather than the health of the turbine as a whole.

6.2 Conclusions

From this report, we are able to draw the following conclusions:

- Strong correlations exist between the temperature of several parts of the turbine. It is possible to create models from these variables, however, for more accurate models we need other variables as well.

- It is possible to identify which generator is being used from the rotor speed. Each generator operates at a distinct speed.

- We are able to apply the methods created by Shewhart to the residuals of a linear regression model to identify out-of-control behavior.

- All models, except the model based on the power output of the starting generator, exhibit the same type of behavior. The behavior that is exhibited, indicates that around the end of 2013, the main generator broke down.

- In the future, these models would probably be able to predict the failure of the turbine.

6.3 Recommendations for further research

This study has shown that condition based prognosis and diagnostics has real potential for wind turbines, and perhaps other (mechanical) applications. In order to improve this concept and improve the practical applicability, further research is suggested. First, we suggest further research into the average run length cycle of the Shewhart Charts for the residuals. Right now, a point is identified as a warning when it lies beyond certain control limits, and many warnings in a short period of time would constitute out-of-control behavior. However, the exact number of warnings that is needed in order for the behavior to be classified as out-of-control is currently unknown. Research into the average run length could provide insight into this, and help construct an objective mathematical standard on what exactly constitutes out-of-control behavior in this context, further eliminating subjectivity from the process. Second, more research could be done
into the more specific identification of the root cause of failures. From the maintenance logs, we knew the main generator malfunctioned, so it was possible to look for behavior that reflected this. If data is available from other wind turbines where other parts than the generator broke down, this might provide insight into the root causes and perhaps allow a precise identification.

In conclusion, we believe that condition based monitoring provides a viable way to monitor wind turbine health, and the application could save a great amount of money due to the timely identification of failures. After all a prevention is better than a cure and in this case time, quite literally, is money.
Bibliography


