MASTER EN INGENIERIA DE SISTEMAS Y CONTROL

PROBABILISTIC APPROACH FOR EARLY FAILURE DETECTION IN WIND TURBINES

MASTER'S FINAL PROJECT

Author:
Miguel Angel García Vaca

Supervisor:
Matilde Santos Peñas

September 2022

UNIVERSIDAD COMPLUTENSE DE MADRID (UCM)
UNIVERSIDAD NACIONAL A DISTANCIA (UNED)

Título:

PROBABILISTIC APPROACH FOR EARLY FAILURE DETECTION IN WIND TURBINES

Proyecto Fin de Master
Tipo−A: Proyecto específico propuesto por un profesor
Master en Ingeniería de Sistemas y Control

Autor: Miguel Ángel García Vaca

Directora: Matilde Santos Peñas

Autorización

Autorizamos a la Universidad Complutense y a la UNED a difundir y utilizar con fines académicos, no comerciales y mencionando expresamente a sus autores, tanto la memoria de este Trabajo Fin de Máster, como el código, la documentación y/o el prototipo desarrollado.

Firmado: Miguel Angel García Vaca

Firma del alumno

Acknowledgments

I would like to thank all those who have made this project possible. First of all, my gratitude to supervisor Matilde Santos for her quick feedback and her time dedicated. Also to Ravi Kumar Pandit for the database provided.

Secondly, a special thanks to my future wife, Maria, for her invaluable support and help in achieving this goal.

Abstract

Usually, maintenance in any electromechanical system is of the preventive type, with a planning of periodic interventions, or of the corrective type, when faults occur in the operation of the system. However, a third type, predictive maintenance, is a clear improvement on the latter two. Aided by the current state of the art and improvements and innovations in machine learning, predictive maintenance is a powerful method of improving system efficiency.

This improvement is crucial in wind turbines where operation and maintenance (O&M) costs are high due to, for example, inaccessible locations, or when technical downtime must be minimized. This is the case for the system proposed in this project, both onshore and mostly offshore wind turbines.

In this project, the aim is to develop a probabilistic assessment of the turbine performance and thus, detect early on incipient anomalies that could lead to a reduction in performance. Of all the possible anomalies that a turbine can present, our case study will be the yaw misalignment, which generates even at low values considerable power losses of wind energy.

For this purpose, we will rely on the data collected by the Supervisory Control and Data Acquisition (SCADA) system, which constantly monitors various variables of interest of the turbine. The most important of these are the power generated and the wind speed, with which we can generate the power curve, that directly reflects the performance of the turbine.

Therefore, one of the objectives will be to model the power curve in order to establish how it behaves under optimal conditions, i.e., what is the dependence between power and wind in a specific wind turbine. To this end, a Gaussian mixture copulas model (GMCM) will be used.

Once this model is generated, the deviations from this model will be evaluated probabilistically for the real case of the same turbine with yaw misalignment present.

Results show that: 1) the gaussian mixtures copula model is able to fit with high accuracy the shape of the power curve and 2) the proposed probabilistic assessment is

able to early detect a yaw misalignment and implies a improve on other methods researched in bibliography.

Keywords

Predictive maintenance, gaussian mixture copula model, power curve, SCADA, wind turbine, yaw misalignment, probabilistic assessment, anomaly detection.

Resumen

Habitualmente el mantenimiento en cualquier sistema electromecánico es de tipo preventivo, con una planificación de intervenciones periódicas, o bien de tipo correctivo, cuando se presentan averías en el funcionamiento del sistema. Sin embargo, un tercer tipo, el mantenimiento predictivo obtiene evidentes mejoras respecto a estos dos últimos. Ayudado por el actual state of art y las mejoras e innovaciones en machine learning, el mantenimiento predictivo supone un potente método de mejora de eficiencia del sistema.

Esta mejora es crucial en turbinas eólicas donde los costes de operación y mantenimiento (O&M) son elevados debido por ejemplo a la localización poco accesible, o cuando los tiempos de parada técnica deben reducirse al mínimo. Tal es el caso del sistema propuesto en el presente proyecto, turbinas eólicas tanto onshore como sobre todo offshore.

En el presente proyecto se tratará de desarrollar una evaluación probabilística del rendimiento de la turbina y así detectar de forma temprana anomalías incipientes que supongan una reducción en este. De entre todos las posibles anomalías que puede presentar una turbina, nuestro caso de estudio será el yaw misalignment, que genera incluso a valores pequeños pérdidas considerables de potencia en la turbina eólica.

Para ello nos apoyaremos en los datos recopilados por el Supervisory Control and Data adquisition (SCADA) que monitorea de forma constante diversas variables de interés para la turbina. De entre todas ellas las más importantes son la potencia generada y el

viento, con la que podemos construir la curva de potencia, la cual nos refleja de forma directa el rendimiento de la turbina.

Por tanto, uno de los objetivos será modelar la curva de potencia para establecer cuál es el comportamiento de esta en condiciones óptimas, es decir, cual es la dependencia entre potencia y viento para una determinada turbina. A tal fin, se usará un modelo de copulas de mezclas gaussianas (GMCM)

Una vez generado este modelo, se evaluará de forma probabilística las desviaciones respecto a este para el caso real de la misma turbina del modelo con desalineamiento del yaw presente.

Los resultados muestran que: 1) el modelo de copula de mezclas gaussianas es capaz de ajustar de manera precisa la forma de la curva de potencia y 2) el método propuesto de evaluación probabilística es capaz de detectar de manera temprana un desalineamiento de la guiñada e implica una mejora respecto a otros métodos investigados en la bibliografía.

Palabras claves:

Mantenimiento predictivo, modelo de copulas de mezclas gaussianas, curva de potencia, SCADA, turbina eólica, desalineamiento de la guiñada, evaluacion probabilistica, detección de anomalías.

Table of Contents

CHAPTER 1: INTRODUCTION 1

1.1 Motivation 1

1.2 General objective and specific objectives 2

1.3 Structure of the report 3

CHAPTER 2: FOUNDATIONS AND STATE OF THE ART 5

2.1 Physical model of the wind energy 6

2.2 Operational regimes of power curve 7

2.3 OEM power curve 8

2.4 IEC standard power curve 8

2.5 Data driven algorithms approach 9

2.6 State of art 10

CHAPTER 3: METHODOLOGY 13

3.1 Dataset and preprocessing 14

3.2 Correction of wind speed 17

3.3 Case of study: Yaw misalignment. 19

3.4 Power curve model using GMCM 20

3.4.1 Modes of the power curve 22

3.4.2 Practical example 24

3.6 P-values 25

3.7 Statistical inference: Fisher's combined probability test 26

CHAPTER 4: RESULTS AND DISCUSSION 29

4.1 Failure in study 29

4.2 Application of the GMCM 30

4.3 Probabilistic assesment 32

4.4 Comparison with other researchs 34

CHAPTER 5: CONCLUSIONS AND FUTURE WORK 37

List of figures 38

List of acronyms 40

Bibliography and references 41

A. Appendix: code developed 44

CHAPTER 1: INTRODUCTION

1.1 Motivation

Energy is considered one of the strategic sectors of any country. In this sense, one of the fields with the greatest future projection is the field of renewable energies, including wind farms [1]. Due to the growing energy demand of the human population and an increasing environmental awareness, it is essential to have enough clean energy sources to satisfy the current and future energy demand, while reducing the impact on the environment. For this reason, renewable energies have become energy alternatives with a great future, as we can see from their rapid growth in recent years [2].

In particular, onshore and offshore wind farms (figure 1) have emerged as one of the main sources of renewable energy. A notable increase in offshore wind farms has been noted in this area. However, making these infrastructures profitable entails minimizing O&M costs, the magnitude of which is high nowadays. This is due to logistical problems in accessing the infrastructure as temporary windows are limited because of often unfavorable weather and sea conditions, leading to longer downtime. Reducing these costs via early detection is crucial.

Figure 1 Off-shore wind farm [3]

The most frequent types of maintenance used in the industry are corrective and preventive. The disadvantages of corrective maintenance are obvious: the action is carried out when the turbine is already stopped and therefore the O&M costs are high due to the inoperability and material and personnel costs derived from correcting this fault. Preventive maintenance makes up for this disadvantage of turbine inoperability by making

use of interventions with a fixed periodicity to avoid this circumstance. However, these interventions are not based on evidence of premature failure, carrying out unnecessary actions on elements that would not require such action and therefore, increasing the frequency of inoperability due to maintenance jobs. In addition, it increases costs associated with these maintenance tasks.

Predictive maintenance, on the other hand, overcomes all the above disadvantages, anticipating failure prematurely and therefore, avoiding unnecessary actions on the wind turbine, reducing O&M costs and maximizing uptime and performance.

In order to implement this type of maintenance, it is vital to have a model of the optimal behavior of the turbine and thus, evaluate deviations from this. Or in other words, this work is focused on detecting as soon as possible when a wind turbine is operating in suboptimal conditions to avoid further damages and costs.

1.2 General objective and specific objectives

The general objective of this final master project is to model the shape of the power curve of a specific turbine using a data-driven algorithm (in particular, gaussian mixtures copula model) and to assess probabilistically when an anomaly exists, and if this model-based assessment is able to detect it early.

The specific objectives of this project include the following:

- Correcting wind speed as a function of air density, in particular ambient temperature.
- With the data previously filtered of outliers, such as negative power values and other factors unrelated to the failure of the turbine itself, select and classify the data belonging to the healthy state and those caused by a yaw misalignment (unhealthy state).
- With the healthy state data create a GMCM model of the output power of the wind turbine to model the joint distribution function.
- Once we know the model, we can estimate the expected value of the power and its deviation from it for each given value of wind speed.

- Probabilistic assessment point-to-point of unhealthy state to test if it can detect this yaw misalignment.
- Adjust this assessment method to detect the failure as soon as possible.
- Minimize or completely eliminate possible false positives in detection.
- Compare the results with other previously established results in the literature using other power curve modelling techniques and probabilistic evaluation techniques.

1.3 Structure of the report

This project report is structured as follows.

- In *chapter 1*, a short introduction to the project is done. In *point 1.1* the motivation and importance of the developed project is presented, and in *point 1.2* the main objective and the specific objectives of the project are described.
- *Chapter 2* is dedicated to the general and most important concepts of a wind turbine. The definition and general characteristics of a wind turbine are presented, devoting *section 2.2* to the most relevant characteristic for this project: the power curve. *Section 2.1* presents the physical modelling of the turbine. *Sections 2.3, 2.4* are devoted to the first approximations for modelling this curve. In *section 2.5* the best candidate for the modelling is introduced. Also, in *section 2.6* the state of the art about various ways of modelling is presented.
- *Chapter 3* is dedicated to the methodology used and the steps followed. In the different sections the concepts and methods followed during the modelling of the power curve and the evaluation of the proposed anomaly detection are detailed. Also in *section 3.4.1* a quick example is presented to visualize the benefits of the described method.
- *Chapter 4* shows the results obtained with the previous methodology and the case of study of this project. It also compares these results with others previously studied cases in the bibliography.
- In *Chapter 5* some conclusions are drawn and future lines of research that can be followed are presented.

CHAPTER 2: FOUNDATIONS AND STATE OF THE ART

A wind turbine is an electromechanical system capable of generating electrical energy from the kinetic energy of the wind. Although there are different types of wind turbines (vertical axis or horizontal axis), our case of study will be a horizontal turbine whose schematic is summarized in figure 2.

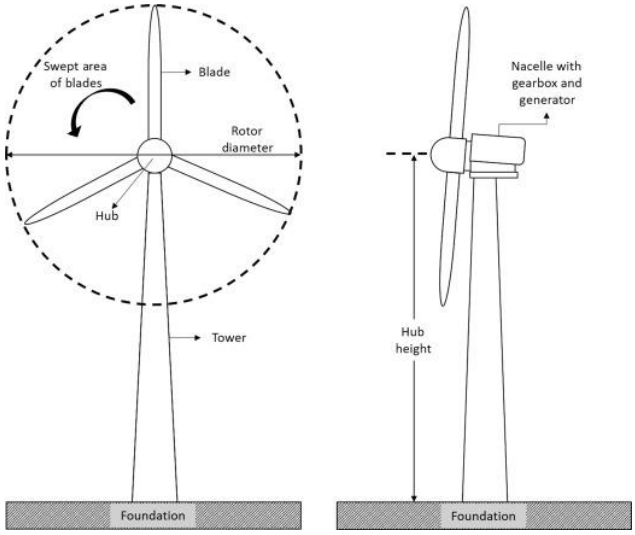


Figure 2 Schematic of a wind turbine [26]

This is the most common and widespread type of wind turbine currently used for large-scale electricity production. They can be located either *onshore* (on land) or *offshore* (on bottom-fixed or floating platforms at sea).

These turbines consist of a vertical tower on which the nacelle is located. In this nacelle there is the electric generator, with the hub where the blades are connected. There are 3 blades arranged vertically; the angle of attack of these blades are controlled to maximize the wind thrust on them. This is called pitch blade control and the turbine that includes this system is called pitch regulated turbine.

We also have a control system that controls the orientation of the nacelle, yaw control, which also serves to maximize wind thrust. As this is our case study, we will go into more detail about this system in section 3.3.

In the nacelle there are different sensors of which the SCADA system is composed. Among them we can find anemometers[1], wind vane, temperature and environmental pressure sensors, as well as many other sensors that are not relevant to the object of this project. The location and technology of these sensors affect their measurements due to, for example, turbulence effects, altitude, hub height, environmental conditions, etc. Since these conditions differ from one turbine to another, the model to be developed will be specific to that particular turbine, and *a priori* cannot be extrapolated to other turbines.

The main characteristic of the wind turbine that interests us is the power curve. This relates the electrical power generated as a function of the existing wind at that moment, and therefore forms the main indicator of the performance of the turbine.

2.1 Physical model of the wind energy

A first approximation of the power curve would be to model the power curve with physical parameters. The equation that relates the dependence between these two variables is:

$$P = 0{,}5\rho A C_p(\lambda, \beta) v^3 \qquad (1)$$

where ρ is air density (kg/m^3), A is swept area by the blades (m^2), C_p is power coefficient which depends on the tip speed ratio λ and pitch blade β (degrees), and v is wind speed at hub height (m/s). Apart from these variables, it should be considered the intensity of turbulence, wind direction, as well as many other factors involved and which influence the functional dependence of the above equation.

Therefore, equation (1) is only a simplified model of the reality that does not reflect the intricate way in which all the factors involved in the power curve are related.

The essential goal is to capture accurately the non-linear dependence of the power curve, as we need to assess in a reliable way the deviations from what it is expected. Consequently, the main problem we face is to develop or choose a method that reflects the specific conditions of a given turbine.

[1] Several anemometers with different measuring technologies are available. https://www.windpowerengineering.com/how-is-wind-measured/

Because of this, the physical model can be discarded as it is not accurate and reliable enough, although it is a simple way to establish the power curve of the system [5].

2.2 Operational regimes of power curve

Three operating regimes can be clearly distinguished in a typical power curve (Figure 3):

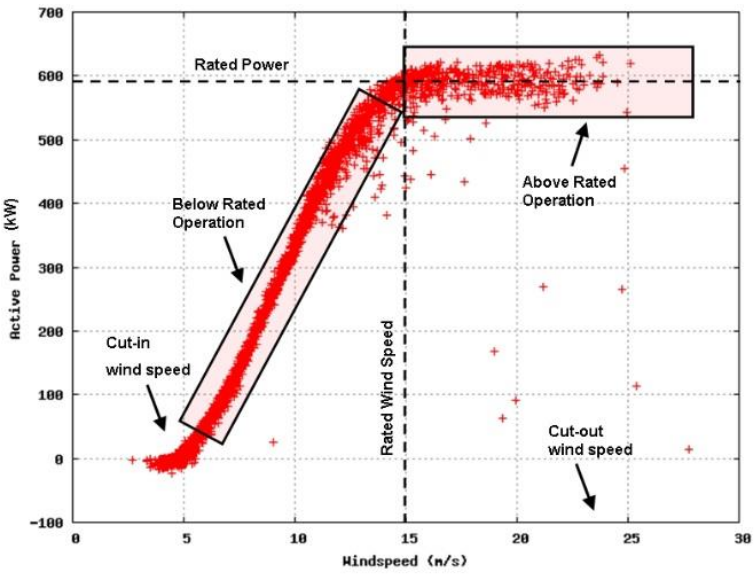


Figure 3 Power curve of wind turbine [4]

- *Cut-in wind speed operation*. When wind speed is below this value, the turbine does not generate power. Around cut-in wind speed, in order to keep the rotor running (or to start it), it is necessary for the electricity network to deliver power towards the turbine, so the power can even be negative.
- *Below rated operation*. In this regime the power delivered is approximately directly proportional to the cube of the wind speed, as it can be seen in the section 2.1. In this regime some different control methods are applied to get the maximum power coefficient for different wind speeds, i.e., to maximize generated power.
- *Rated operation*. From rated wind speed value, the turbine delivers the maximum possible power, and it must be kept constant. The turbine uses different control methods, e.g. the pitch blade angle limits the power delivered up to the rated power value.

- *Above the cut-out wind speed value*, the control system stops the turbine to avoid possible damage.

2.3 OEM power curve

Another approach would be to establish the power curve as the empirical power curve supplied by the manufacturer. However, this curve is created in a wind tunnel under standard conditions using methods that differ from the conditions of the SCADA sensors and far from considering the environmental conditions of the turbine

As proof that the OEM curve is not faithful to the real conditions of the turbine, figure 4 shows the power curve obtained with data and the curve given by manufacturer's specifications. It clearly differs from the actual functional dependence of our turbine. Blue dots are the healthy data of our dataset and red line is the fit of the manufacturer data.

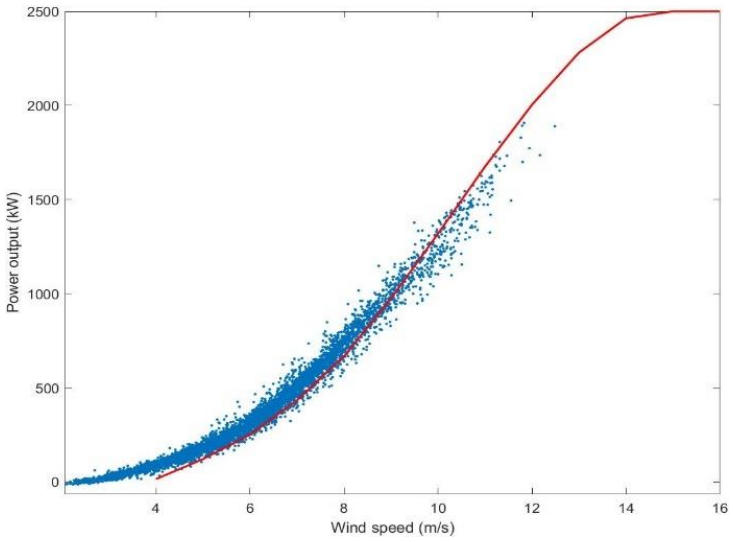


Figure 4 OEM power curve vs dataset

2.4 IEC standard power curve

Another way to obtain the power curve is to use the international standard, in our case, IEC 61400-12-1:2017 [6]. This develops the internationally accepted way of generating the power curve of a wind turbine. In a simplified way, we must divide and group the dataset into bins of 0.5 m/s, and then calculate the mean and standard deviation

of the power output for each bin. In figure 5 is represented by blue dots our dataset. The power mean values of each bin are connected by the red line, and the error bars in red represent the standard deviation for each bin.

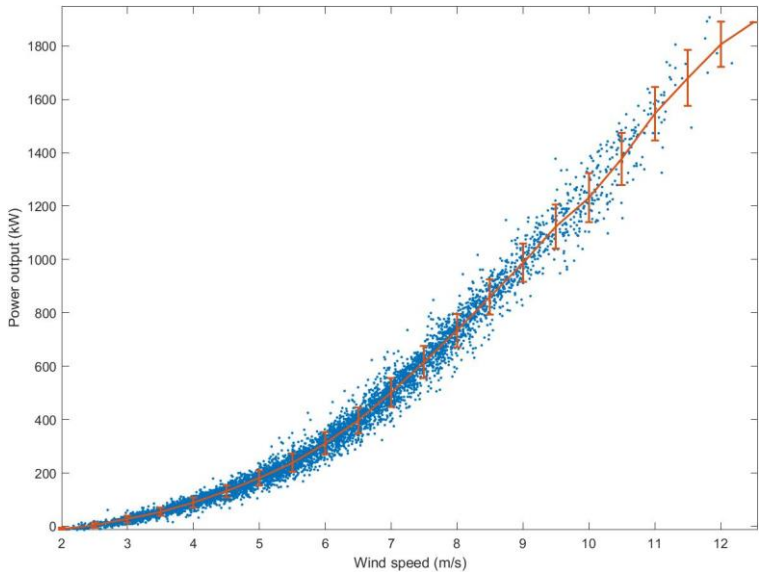


Figure 5 IEC standard power curve

However, the power curve generated in this way is not accurate due to the width of the intervals used (0.5 m/s) and the rapid growth of the power due to the cubic nature with respect to the wind speed. In conclusion, it is only a simple approximation.

2.5 Data driven algorithms approach

From the above, it can be concluded that data driven algorithms are one of best candidates to capture the shape of the power curve that best fits reality, i.e. the specific conditions of our turbine.

In addition, another problem we face is to discern when this deviation from the optimum is due to a failure in the turbine[2]; or if this deviation is caused by faults in the data acquisition system due to flaws with: measurement sensors, communication with SCADA system, the data acquisition method itself, random data noise, or any other faults not attributable to a failure in the turbine. To solve this problem, we will first eliminate all

[2] in this project we will focus on yaw misalignment.

outliers from our reference model (more details in section 3.1), as well as establish a probabilistic assessment of the failures and thus, eliminate possible false positives.

2.6 State of art

Nowadays there is a multitude of publications about the mentioned issue of determining the shape of the wind turbine power curve in a reliable and trustworthy way. There is not a clear winner approach among the authors. Among all of them, those that focus on data-driven algorithms could be highlighted as follows:

- *Wind Turbine Condition Assessment through Power Curve Copula Modeling.* Gill, Stephen, Galloway [7]. The purpose of this paper is to develop a probabilistic model to estimate the distribution function of the power curve, in order to detect deviations from the expected behavior. For the probabilistic model, empirical copulas are proposed. The operational data used come from a 5-min averaged SCADA system of two onshore fault free turbines located in Scotland. Possible anomalies like blade faults, yaw system faults or pitch system faults and how they influence differently in empirical copulas are studied.
- *Power Curve Based Online Condition Monitoring for Wind Turbines*. Wang, Infield [9]. This paper proposes a copula-based method to improve the expected power curve based on the IEC standard. The copulas method is used for the rejection of existing outliers in the dataset. With this improved power curve, a method is established to evaluate the statistically significant deviations in the case study of yaw misalignment. The operational data used are from a 10-min averaged SCADA system of a fault-free on-shore wind turbine.
- *Gaussian Process Operational Curves for Wind Turbine Condition Monitoring*. Pandit, Infield [11]. In this paper a Gaussian Process method is proposed to estimate the reference power curve, the pitch angle curve as a function of wind speed and rotor speed as a function of wind speed. Gaussian process is a Bayesian, non-parametric method for non-linear regressions. The operational data used are from SCADA data 10-min averaged from an onshore wind turbine.

- *Condition monitoring of the power output of wind turbine generators using wavelets*. Watson, Xiang, Yang, Tavner [12]. In this paper the power output of a variable-speed wind turbine generator is monitored, and this signal is processed with continuous wavelet transforms to find the frequency components that are characteristic of failure. The case of study are two wind turbines 1.5MW variable-speed pitch-regulated where a misalignment of the high-speed shaft has been detected.
- *Online wind turbine fault detection through automated SCADA data analysis*. Zaher, McArthur, Infield, Patel [13]. This paper proposes a multi-agent system that monitors and analyses the different SCADA variables of a wind turbine detecting abnormal behavior of these variables that may indicate an incipient failure. For the modelling of normal behavior, it uses multilayer neural networks in certain SCADA variables of interest such as: active power output, wind speed, nacelle temperature, power factor or gearbox oil temperature.
- *Monitoring of wind farms' power curves using machine learning techniques*. Marvuglia, Messineo [14]. In this paper three data driven approaches are proposed to estimate the power curve relationship: self-supervised neural network called GMR (Generalized Mapping Regressor), a feed-forward Multi-Layer Perceptron (MLP) and a General Regression Neural Network (GRNN). The data used comes from an hourly averaged SCADA system for an entire year and an entire wind farm. To monitor deviations between predicted and observed power, a residual control chart technique is used.
- *Wind turbine multi-fault detection and classification based on SCADA data*. Vidal, Pozo, Tutivén [15]. The wind turbine under study is simulated SCADA data of a 4,8 MW three-blade horizontal variable speed. In this simulation different faults scenarios are recreated related to the pitch blade system. Besides, it is only simulated the operational regime above rated operation. It is proposed to use multi-class support vector machines classifier to detect these faults. Before training the model, the different SCADA signals are pre-processed and the most relevant ones are selected by a principal component analysis.

CHAPTER 3: METHODOLOGY

As a brief introduction of the methodology that has been followed in this project, the flowchart in figure 6 summarizes the main steps.

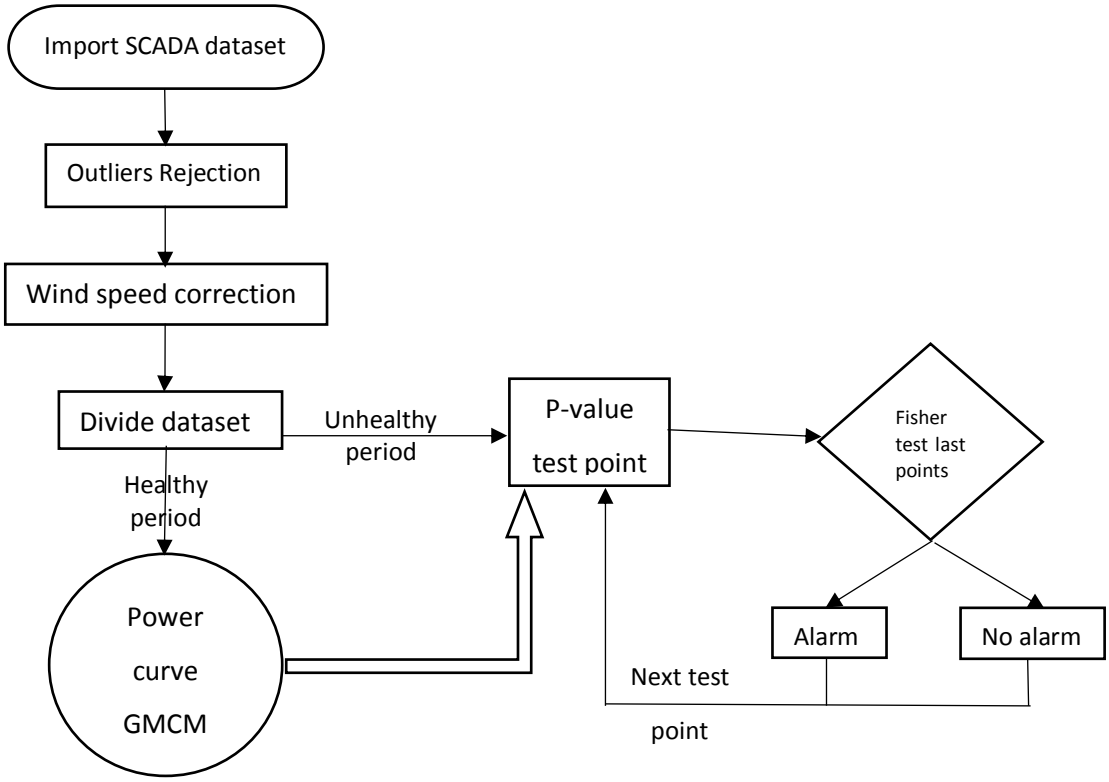


Figure 6 Flowchart general methodology

A short explanation of each step is as follows, although details on each step will be given in the following sections.

- The first step is to load the dataset with all the data that will be needed, in addition to the rejection of the outliers, if applicable. Besides, the relative yaw error is used to obtain the absolute yaw error.
- The wind correction will then be made according to the air density, which depends on the atmospheric pressure and temperature.
- Next step is to divide the dataset into data without yaw misalignment present (healthy period) and data with yaw misalignment (unhealthy period).

- With the healthy period data (optimal turbine conditions), the power curve modelling will be carried out using the proposed GMCM.
- At each point of the unhealthy period, and using the previously created model, we will calculate the p-value of the measured electric power based on the probability distribution at that point according to our model.
- To adjust the efficiency of the anomaly detection, several consecutive p-values will be combined using the Fisher test[3]. This will tell us whether an alarm is triggered at the analyzed point or not.
- Calculate the p-value of the next time point and perform Fisher's test and so on for all points in the time series.

3.1 Dataset and preprocessing

The dataset used comes from an operational variable pitch regulated turbine manufactured by Siemens with a rated power of 2.5 MW. This turbine has a built-in SCADA system that collects data for more than 100 different variables such as power, wind speed, temperature, rotor speed and so on. These variables are collected every 10 min and the statistical values of mean value, maximum value, minimum value or standard deviation are shown.

In our case, we have a partial extract of these SCADA variables whose sample can be seen in figure 7. The summary of these variables is:

tstamp	datetimestamp	WindSpeed_AVG	LVActivePower_AVG	EnvAirTempOutside_AVG	YawError
1238546400	2009-04-01 00:40:00.000	4,102055	69,78228	7,056631	9,3016
1238547000	2009-04-01 00:50:00.000	5,032456	173,0703	7,074712	-3,0029
1238583600	2009-04-01 11:00:00.000	3,593298	19,28885	10,59632	0,1747
1238586000	2009-04-01 11:40:00.000	4,060245	32,88626	11,37038	-2,0686
1238586600	2009-04-01 11:50:00.000	3,6722	15,87732	11,5101	-2,775

Figure 7 Dataset used

- ***Datetimestamp*** (*Year-Month-Day Hours:Minutes*) Time stamp.

[3] The number of points needed will be adjusted based on our data to minimize false positives.

- **WindSpeedAVG** *(m/s)* Mean value of wind speed measured by anemometer at the nacelle.
- ***LvActiverPowerAVG*** (*kW)* Mean value of the power delivered by the turbine
- ***EnvAirTempOutsideAVG*** (*ºC*) Mean value of ambient temperature.
- ***YawError*** (*degrees)* Relative value of yaw misalignment regarding to the wind direction.

These data cover a temporary window from 1/04/2009 to 30/06/2009. In total there are 5829 points between healthy data and unhealthy data with yaw misalignment. This unhealthy data are 200 points.

The SCADA data from which the used dataset comes from is not free of outliers, so it has been previously cleaned. This outliers rejection of our dataset has already been carried out in previous works by other authors following the methodology described in [17].

These outliers are due to different reasons, but in no case are indicative of failure of any of the turbine subsystems. A complete and detailed study of the outliers and its causes in a wind turbine can be found in reference [28]. As a resume, among them the following cases can be highlighted:

- Turbine operational transitions. Since data are captured every 10 min, it is possible that in this time interval the turbine has changed its operational state (on-off) causing a point too far away from normal operation. These outliers are characterized by the fact that they are isolated and do not show any pattern (Figure 8, region named as B)
- Errors or noise in the sensors or sensor communication with SCADA. (Figure 8, region named as B)
- Lost data.
- Negative powers due to being below the cut-in speed. (Figure 8, region named as A)
- Mismatched timestamps.

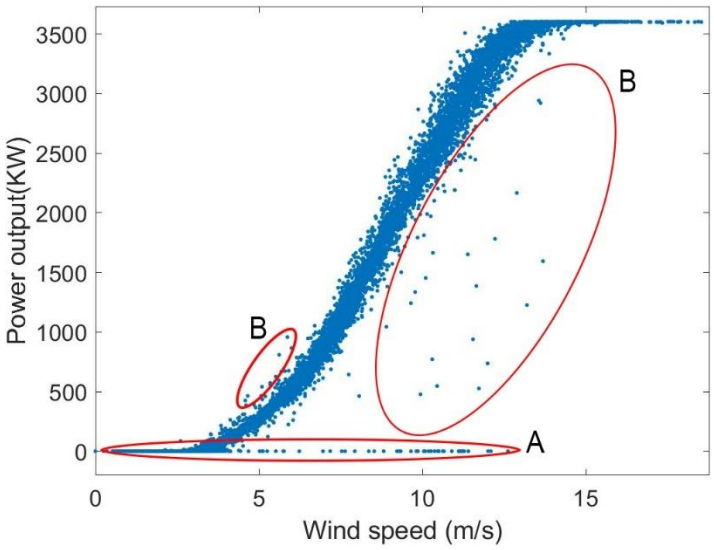


Figure 8 Example of outliers in a power curve

As an example, in figure 8 it can be seen the raw data of two months from SCADA of Yalova's wind farm dataset[4] available on Kaggle. This consists in SCADA data for 10-minute averaging interval, similar to the paper which it is based this project.

There are a multitude of methods for outliers rejection, which is considered a field of study in its own right. As an example, we will proceed to use GMCM as a method to reject the outliers present in figure 8. The case of type A outliers is simpler, it is only necessary to define a rule that eliminates all the powers equal to or less than 0.

The GMCM model is created for the example of this turbine, the result of which can be seen in figure 16. Since this is one of the specific objectives of the project, more details on GMCM modelling will be given in section 3.4.

Next step is select a contour level in this model to reject all the outliers that are outside this contour and far enough of the power curve as isolated points. For that purpose, based on methodology in reference [23], it is chosen three standard deviations in the bin closest to the rated wind speed (12.5 m/s). The reason for the choice of 3 σ is to capture not only the points belonging to the healthy state of the turbine but also those belonging to possible anomalies of the system (such as the case of yaw misalignment). The result can be seen in figure 9, where black dots are raw data and red line is the chosen contour level.

[4] https://www.kaggle.com/berkerisen/wind-turbine-scada-dataset

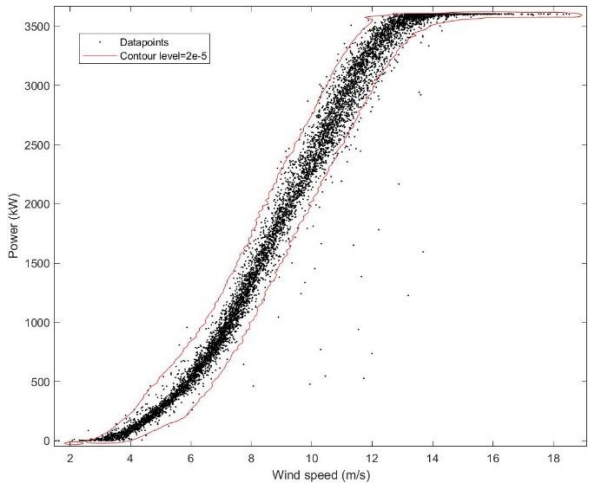


Figure 9 Contour level of GMCM for outliers detection purpose

The result of the cleaning of the outliers of the figure 8, whose SCADA data are extracted from Yalova's wind farm dataset, can be seen in figure 10.

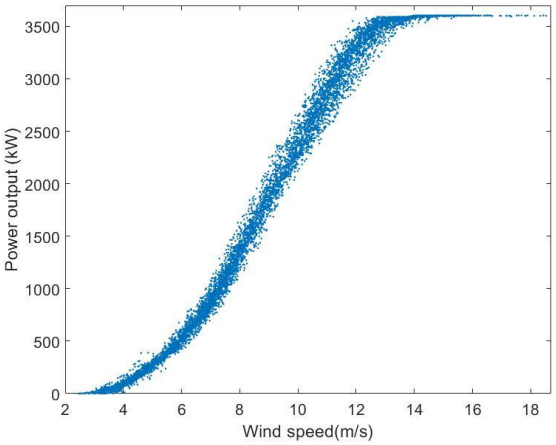


Figure 10 Power curve after outliers rejection

3.2 Correction of wind speed

As seen in section 2.1 equation (1), the power depends cubically on the wind speed and wind speed depends on the atmospheric conditions, so it is necessary that to have a wind speed model as accurate as possible. In particular, wind speed depends on the density of the air as shown in the following equation:

$$V_c = V_m \left(\frac{\rho}{1.225}\right)^{\frac{1}{3}} \quad (2)$$

where V_c is corrected wind speed(m/s), V_m is the measured wind speed(m/s), and ρ is the density of the air (kg/m^3), that can be considered as:

$$\rho = 1.225 \left(\frac{288.15}{T}\right)\left(\frac{B}{1013.5}\right) \quad (3)$$

where T is the ambient temperature (Kelvin) and B the atmospheric pressure (mbar).

From this equation and the SCADA ambient temperature data for each point that we have, the wind speed of the dataset is corrected [5]. Some samples of these corrections are shown in figure 11, where it can be seen corrected wind speed (V_c), measured wind speed (V_m), ambient temperature and relative error, successively.

V_c(m/s))	*V_m(m/s))*	*EnvAirTempOutside_AVG(ºC)*	*Relative Error*
9,473689561	9,370124	5,652838	0,010932
3,179619334	3,141864	4,856777	0,011874
3,168819325	3,15796	12,04773	0,003427
2,733489461	2,74972	20,16335	-0,00594
2,609748544	2,633635	22,98476	-0,00915
7,457090708	7,354369	3,255446	0,013775

Figure 11 Samples of wind speed corrected

Last two rows represent the minimum and maximum relative errors for the entire dataset. The minimum relative error would have a value of -0.00915 with an associated absolute error of -0.024 m/s for a temperature of 22.98ºC, and the maximum relative error would be 0.013775 with an absolute error of 0.103 m/s for a temperature of 3.25ºC.

From equation (2) and (3) it can be seen that the correction equal to 0 would occur at an ambient temperature of 15ºC, the further we move away from this temperature both above and below, the higher this correction will be. Likewise, the higher the measured wind speed, the higher the value of the correction will be.

It can be concluded that, based on the magnitude of the absolute errors of the wind speed and given the cubic dependence of this wind speed for the power generated by the turbine, this correction is important in order to obtain the most reliable wind speed measurement possible regardless of the atmospheric conditions.

[5] Atmospheric pressure data is not available, so it is considered constant and equal to 1.

3.3 Case of study: Yaw misalignment.

Of all the possible anomalies that can be present in a turbine and generate a loss of performance, we will focus on the yaw misalignment for which we have available data.

One of the ways to maximize the efficiency of a wind turbine is to control the yaw. The best way to take advantage of the available wind is facing the turbine towards the wind direction. This is achieved by controlling the orientation of the turbine nacelle in real time, monitoring the wind direction at all times so that any changes cause an adjustment in the angle of the nacelle. This is called yaw control.

Figure 12 shows the different elements of a turbine and the yaw control system. Inside the nacelle there is a low speed shift shaft, connected to a gearbox that increases the rotational speed of the secondary shaft, high speed shaft, where the electric generator is located. It can also be seen how each blade can vary its angle, which is called pitch. Regarding the yaw control, at the base of the nacelle is located the yaw motor and the yaw drive, which is in charge of turning the nacelle to direct it in the direction of the wind. The orders come from a controller located in the nacelle that receives the data of wind direction coming from the wind vane located on the outside of the nacelle.

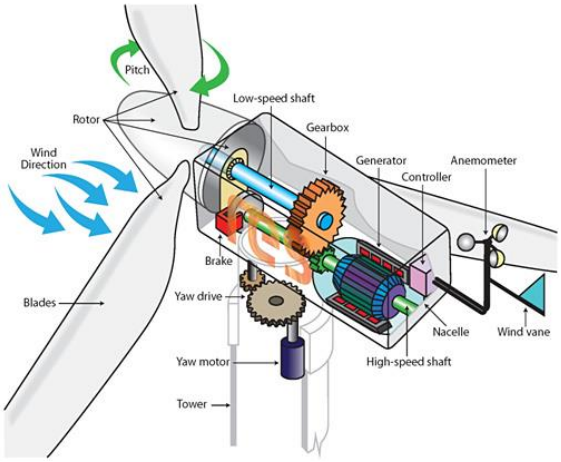


Figure 12 Elements of yaw control system [18]

A failure in this control system, vane misalignment, or any other issue can lead to an error in the yaw, which results in the turbine operating under suboptimal conditions. Several studies and empirical experiments show that the power loss depends on the square of the cosine of the yaw misalignment [27], as it can be seen in equation (4):

$$P_{loss} = P_{rated}(1 - (cos\Delta\varphi)^2 \quad (4)$$

where P_{loss} is the power loss (kW), P_{rated} is the rated power of the turbine (KW) and $\Delta\varphi$ is the yaw error.

This means that even error of 10° means a loss of around 3% efficiency [19]. It is the main purpose of this project to detect as soon as possible when this circumstance occurs.

From the data of the wind direction measured by the nacelle vane and the yaw direction extracted from the tracking system, we can calculate the yaw error, which is one of the variables we have. In figure 7 it can be seen that yaw error in the dataset is positive or negative. This is because the yaw error is measured against the nacelle orientation. In other words, the error is not absolute, so first step is to calculate the absolute value of the yaw error.

Yaw error is a value that must range from -180º to 180º. Therefore, in the dataset is necessary to correct when the measured yaw error is out of this range, converting it inside the accepted range by adding one turn (360º).

3.4 Power curve model using GMCM

For modeling the power curve, the Gaussian mixture copulas model (GMCM) will be used, which is a data-driven method for unsupervised clustering. This method is based on statistical copulas which are a very useful tool in the case of dependent variables with a non-linear correlation. The advantage of copulas is that they are able to separate the marginal distributions from the dependence between them.

According to Sklar's theorem [20], the joint probability distribution function (PDF) of 2 variables follows the equation:

$$f(x_1, x_2) = c(u_1, u_2) f_1(x_1) f_2(x_2) \qquad (5)$$

where f_i denotes the marginal distribution PDF of variable *i,* u_i denotes marginal CDF of variable *i*, and *c* is the copula density function relating the marginal CDF.

There are multiple families of copulas in mathematical theory such as Frank Copula, Clayton, Archimedes and others [16]. The goal is to find a copula that captures the

dependence between CDFs in multimodal distributions[6]. But none of the above fulfils this objective.

One candidate would be Gaussian mixtures model (GMM), in which it is assumed that the different modes are normally distributed and therefore they form elliptically shaped modes. However, this assumption, that the clustering is elliptical, is not always fulfilled, as in our case. To alleviate this problem, GMCM is used, whose main ability is to characterize multimodal distributions in which some of these modes are non-elliptic. An example of these benefits in these kinds of distributions can be seen in figure 13 and figure 14.

Figure 13 upper left corner shows the scatter plot of a synthetic bivariate data, in the upper right and lower corner the marginal distributions of both dimensions, in red line the best fit for these marginals is represented. The synthetic bivariate data correspond to three modes (or clusters). The mode centered at x=1, y=0 has an elliptic shape and the other two are non-elliptic (non-Gaussian) with a strongly non-linear correlation between x and y.

Figure 14 (a) and (b) plots the contour lines of the joint PDF using GMCM and GMM, respectively. The GMM results in only elliptic (Gaussian) modes that poorly model the underlying distribution. However, the GMCM results in better modelling by capturing not only the elliptic mode but also the non-elliptic modes where the dependence is not linear. This results in a better accuracy of the GMCM.

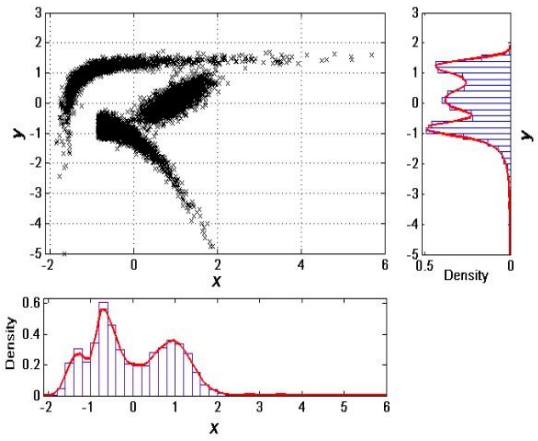


Figure 13 Synthetic data for visualization of GMCM benefits [21]

[6] The mode of a distribution is the value at which the joint PDF has a local maximum.

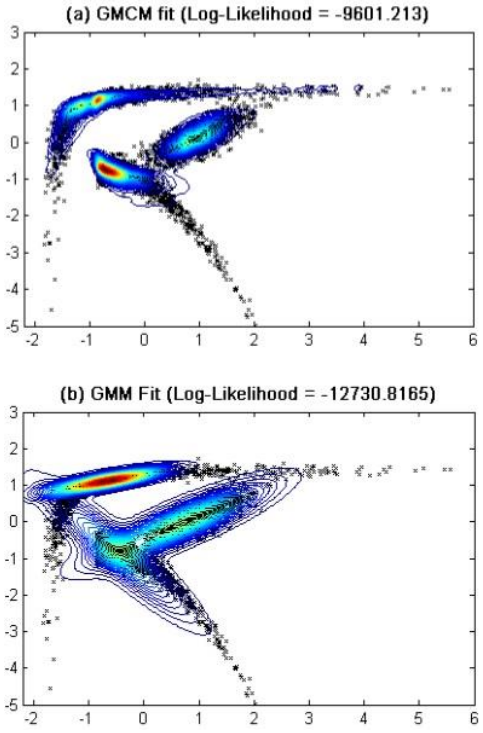


Figure 14 GMM and GMCM fitting of figure 13 [21]

The GMCM for bivariate data follows the equation below [21],

$$c(u_1, u_2; \theta) = \frac{\varphi(\phi_1^{-1}(u_1), \phi_2^{-1}(u_2); \theta)}{\varphi_1\left(\phi_1^{-1}(u_1)\right) \varphi_2(\phi_2^{-1}(u_2))} \quad (6)$$

where φ_i denotes marginal density and ϕ_i^{-1} inverse marginal distribution. θ is the set of parameters consisting of the mean and covariance matrices of all modes, as well as the mixing ratios of each mode. In total, there will be 18 parameters.

For the computation of this GMCM, the code in reference [22] has been used. This code is developed by the authors in MATLAB software.

3.4.1 Modes of the power curve

In reference [23], it is used SOM to determine the number of modes present in a power curve. The result is the number of modes of the power curve is 3. Each of them corresponds to the operational regimes of the wind turbine seen previously.

A self-organizing map, SOM, is a type of artificial neural network that produces a low-dimensional representation of a high-dimensional dataset. In our case, it is able to capture the numbers of clusters which make up the power curve of a wind turbine.

In the same reference, the example of figure 15 is attached to visualize the goodness of the fitting of the GMCM for power curves compared to other options.

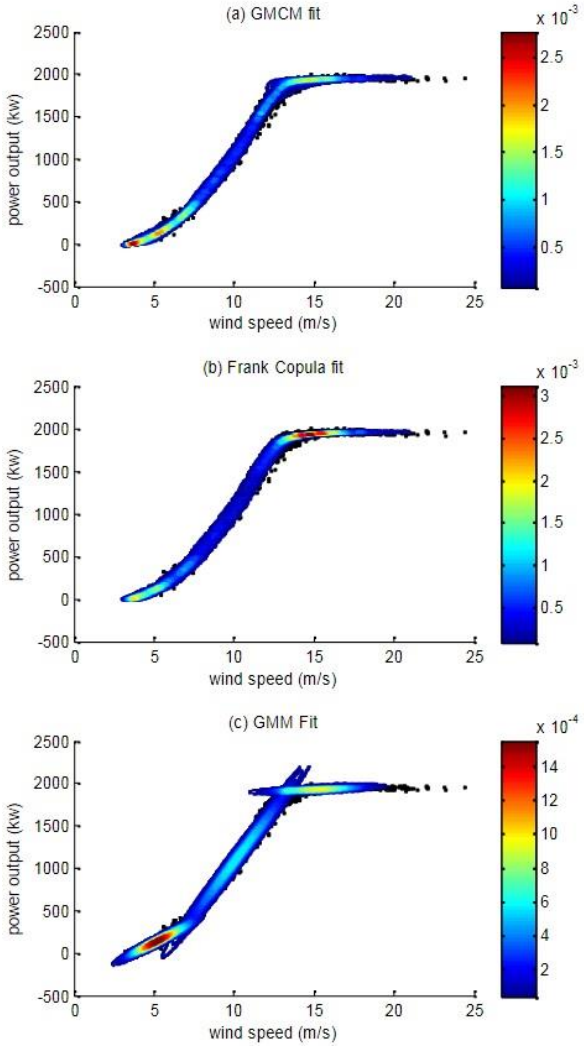


Figure 15 Comparison between types of copulas [23]

As can be seen, the GMCM has clear advantages over other options such as the Copula Frank or GMM, the most important of which is that it more accurately matches the different modes of the power curve, making it a more precise model.

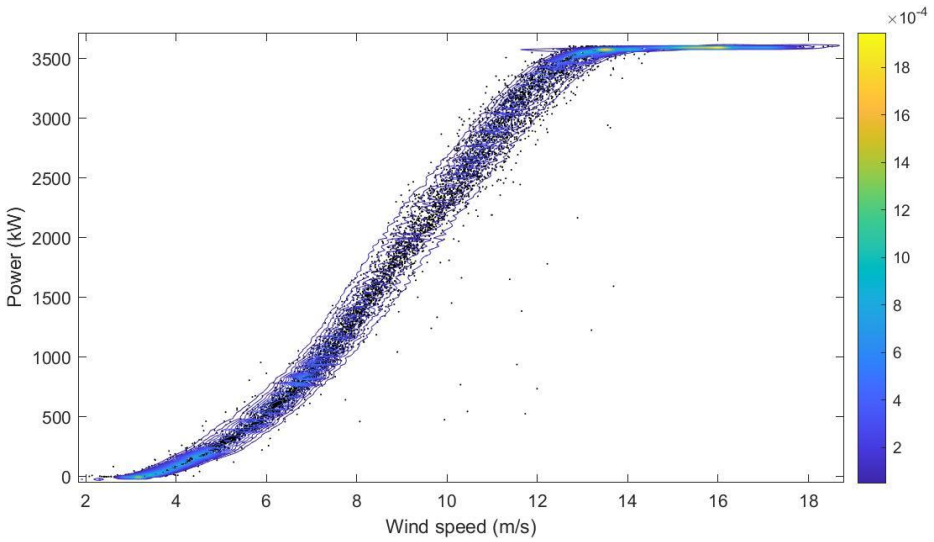
3.4.2 Practical example

As a quick example with our own data, GMCM will be performed for Yalova's wind farm dataset[7] available on Kaggle. This dataset, like the one used for the main objective of the project, uses SCADA data for 10-minute intervals.

The first steps of the methodology are developed. First step was the outliers rejection in the dataset, that includes the removing of all power points that are smaller than or equal to zero or undetermined values. Dataset consists of two months (August and September). After that, there are 7646 valid points.

The following step was the correction of wind speed in relation to air density. Due to the absence of the air conditions in the SCADA dataset, it is not possible to correct the wind data according to the standard air conditions, so for this example, these corrections are not applied.

Now we apply the technique to model the data. It obtains the following joint density distribution (top) and the marginal PDF (bottom, left and right) for the power and wind speed variables (figure 16).



[7] https://www.kaggle.com/berkerisen/wind-turbine-scada-dataset

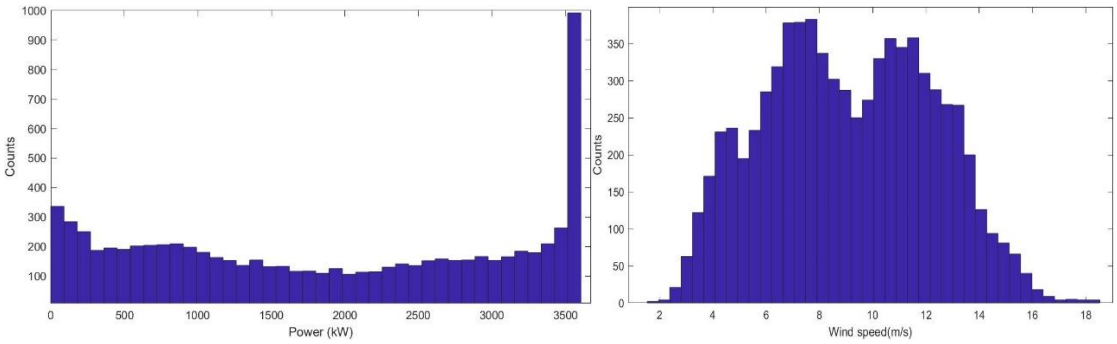


Figure 16 GMCM fit and CDF for Yalova example

As we can see, the GMCM fits well to the data following the shape of the power curve and the three modes that it is possible to observe in the marginal PDF's.

3.6 P-values

In the present project, the unhealthy state of the dataset will be evaluated point by point, that is, if these can be probabilistically considered a failure or not. In other words, we must compare the measured power against the expected power for that given value of wind speed and its associated uncertainty. To account for this last point, expected value and its uncertainty, we will calculate the p-value of the observed power, which will be obtained on the basis of the GMCM model obtained previously. More details as follows.

As output of the GMCM model, we have the value of the joint probability distribution function, denoted from now on as Z, calculated on a grid of points which we denote as ($WindSpeed_{model}$,$Power_{model}$)[8]. Likewise, the point at which we want to calculate its p-value will be denoted as ($WindSpeed_{test}$, $Power_{test}$). The objective of this section is to calculate, fixed the wind speed at $WindSpeed_{test}$, the probability that power output is less than observed power ($Power_{test}$), i.e:

$$P\text{-}value(WindSpeed_{test}, Power_{test}) = P(Power < Power_{test} \mid WindSpeed = WindSpeed_{test})$$

This quantity is calculated mathematically according to the following equation:

$$Pvalue = \frac{\int_0^{Power_{test}} Z(Power, WindSpeed_{test})\, dPower}{\int_0^{\infty} Z(Power, WindSpeed_{test})\, dPower} \qquad (7)$$

[8] In the developed code is denoted by (x_{grid}, y_{grid}, Z)

It will perform the procedure for every point of the unhealthy period. As output the *Pvalue* for each point is obtained. So first we have the model grid ($WindSpeed_{model}$,$Power_{model}$) for the $WindSpeed_{model}$ value closest to our test point, $WindSpeed_{test}$. Once this is obtained, we have the function $Z(Power, WindSpeed_{test})$ with which to calculate the P-value.

As an example, in figure 17 it can be seen this Z function, which is a PDF, for a wind speed value of 4 m/s, according to our GMCM built with 1024 grid points. The p-value will be the area enclosed by the function between the observed power ($Power_{test}$) and power equal to 0.

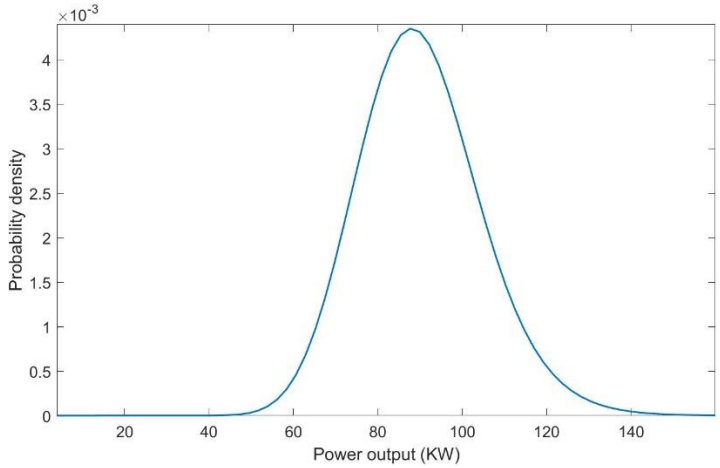


Figure 17 Example of Z at wind speed of 4m/s

3.7 Statistical inference: Fisher's combined probability test

The above procedure described in subsection 3.6 only takes into account a single test point, however this is not sufficient to establish strong evidence without making errors. If we consider a set of *k* independent tests, all these obtained p-values can be combined in a joint test to determine if the null hypothesis is fulfilled.

In our case, the null hypothesis H_0 is that there is no association between the two variables (power and wind) due to randomness. Therefore, if the test value is lower than the significance level (or confidence interval), it can be inferred that it is not due to any randomness and it is due to a physical anomaly.

There are several ways to combine independent tests, among them is the Fisher's combined probability test [24], which is the one that will be used. It is based on the fact that several non-significant results, if they occur consecutively, may suggest significance, i.e. even if one test point is not too far away from the expected value of the power, if we consecutively combine it with the following points (combining their p-values), they may indicate an anomaly.

Fisher's test is a non-parametric statistical test, similar to chi-square whose main advantage, and the reason for its use here, is that it is valid for small statistical samples. The equation for the test is:

$$\chi^2_{2k} \sim -2 \sum_{i=1}^{k} ln\,(pvalue_i) \tag{8}$$

where $pvalue_i$ is the probability of test i that it was calculated previously. That is, it follows a chi-square distribution with 2k degrees of freedom.

It can be concluded that if the value of the Fisher's test is lower than the significance level, the wind turbine has a failure. It is among the objectives of this work to adjust how many k tests are necessary and the significance level for the detection of anomalies, minimizing false positives or false negatives.

CHAPTER 4: RESULTS AND DISCUSSION

This chapter presents the results of applying GMCM on the power curve of an onshore variable pitch regulated wind turbine manufactured by Siemens with a rated power of 2.5 MW. Likewise, it is presented the results of the probabilistic evaluation on our case of study of a yaw misalignment confirmed by other techniques. Finally, the results obtained will be compared with other studies using different methods carried out on the same turbine and with the same yaw misalignment.

4.1 Failure in study

In figure 18 the absolute yaw error of the entire dataset is represented as a time series. Visually, it can be seen how in one zone the yaw error shoots up to 50º, which indicates a more than probable failure of the yaw control (unhealthy period). This zone corresponds to the time window from 14/04/2009 at 11:20 to 18/04/2009 at 14:00 (figure 19). As indicated above, this consists of 200 points, which includes both the unhealthy data as well as a pre- and post- unhealthy fraction data for visualization and adjustment of possible false positives.

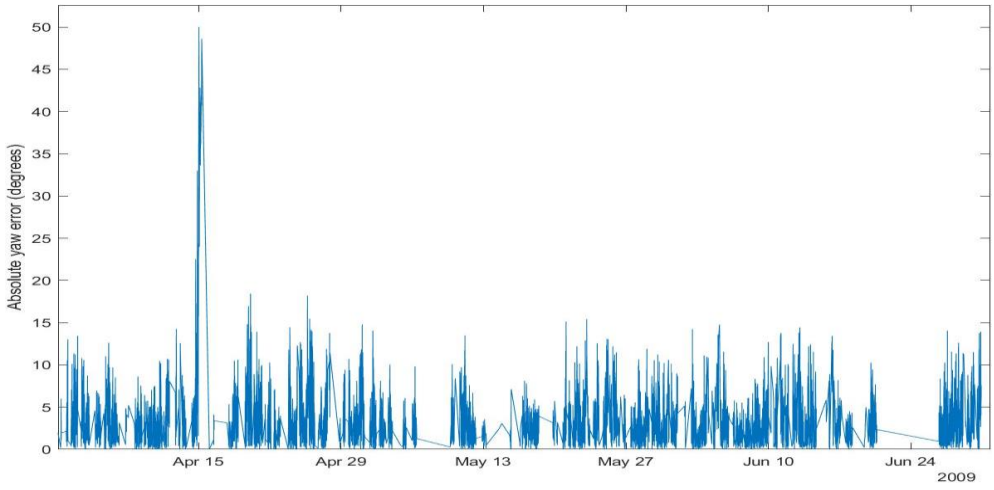


Figure 18 Absolute yaw error vs time

Zooming in on the above, the unhealthy period shown in figure 19 will be tested with the application of the detection of the yaw misalignment.

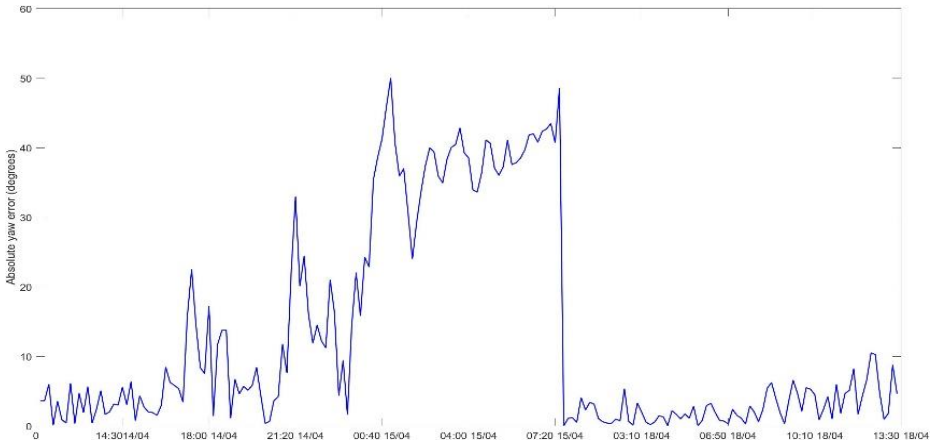


Figure 19 Unhealthy period

As a confirmation of the existence of a yaw misalignment in this time window, from [25], which uses the same dataset, we can observe in figure 20 that from a certain instant the direction of the nacelle becomes stuck. That instant of time corresponds to 17:00 on, 14/04/2009.

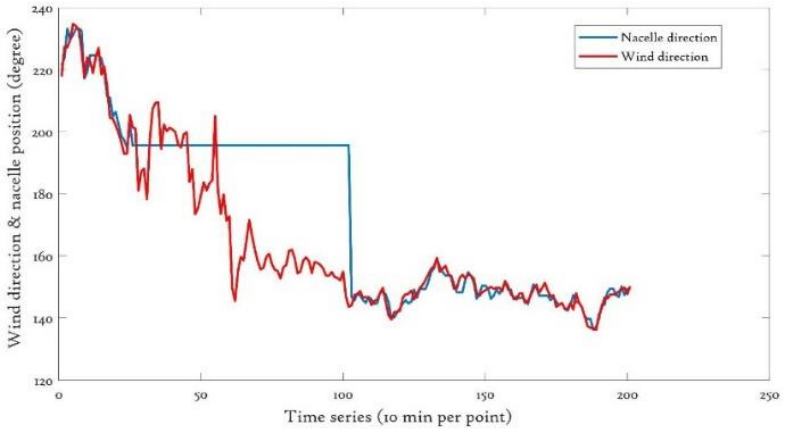


Figure 20 Wind & nacelle direction unhealthy period [25]

4.2 Application of the GMCM

Now, the GMCM is generated with the healthy data of the wind turbine. The input parameters for the model are the following:

- *Numbers of modes, K=2*. Since dataset do not have the region above the rated power, the number of modes (or clusters) will be equal to 2.
- *Numbers of bins=2*^{12} This is the number of bins used to estimate marginal densities (PDF).

- *Numbers of iterations=200*. This is the maximum number of iterations used by the algorithm to estimate the GMCM. This number of iterations is never exceeded, being only 60 iterations necessary in any case.
- *Number of grid points*. It is the number of points of power output and wind speed where the joint PDF is calculated. It is set at 1024x1024 points.

Figure 21 shows the resulting contour lines of the GMCM. On the other hand, figure 22 shows the dataset in the calculated copula space and in red line the function y=x, which expresses the dependence between the marginal distributions of power and wind speed.

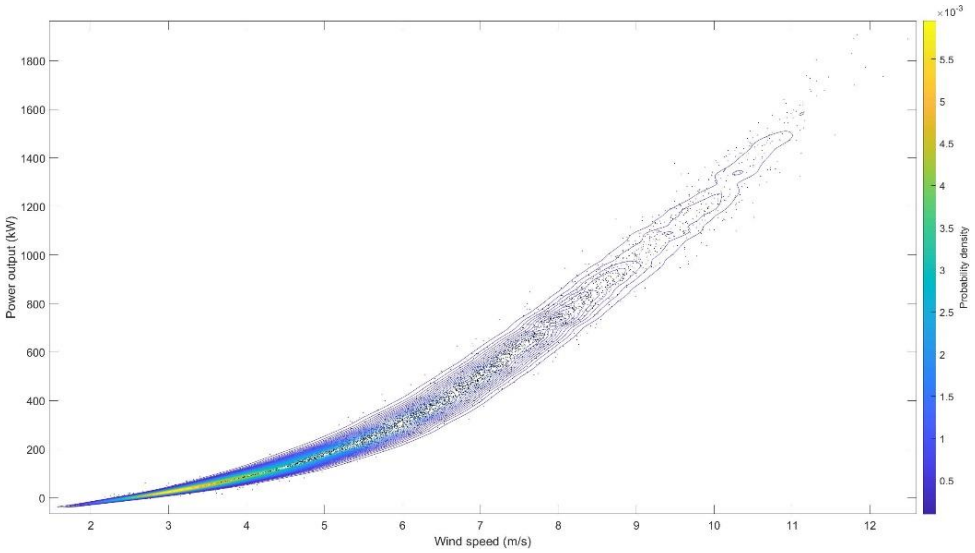


Figure 21 GMCM fit of healthy period

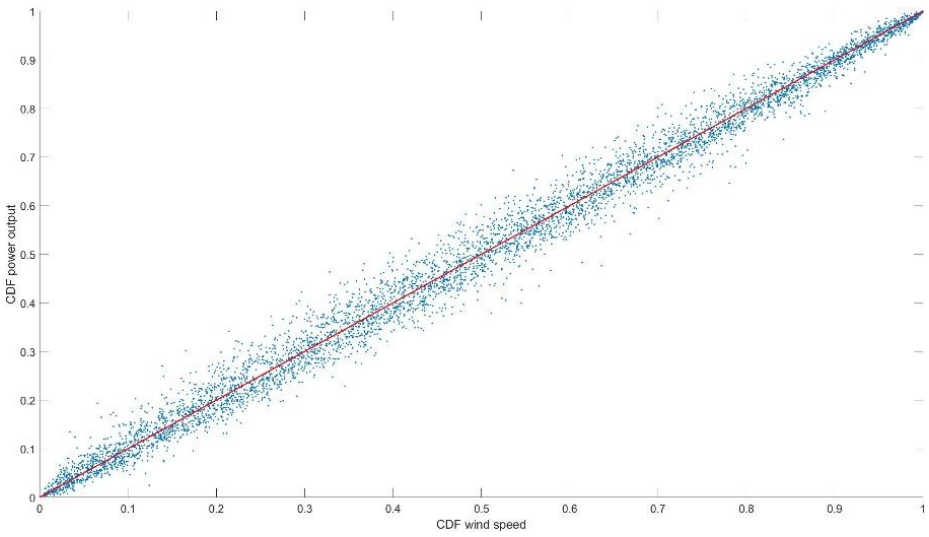


Figure 22 Dataset into copula space

It can be seen that the GMCM fits the data very closely. Especially noteworthy is the smooth coupling between the two modes of the power curve, as we also saw in figures 15 and 16, which is one of the strengths of the GMCM.

To visualize the goodness of the fitting of the GMCM, figure 23 plots the histogram of frequencies of the residuals, considering the residual as the difference between the expected value of power for each point and the observed power. The Gaussian distribution with confidence bound of 95% is also drawn in red line.

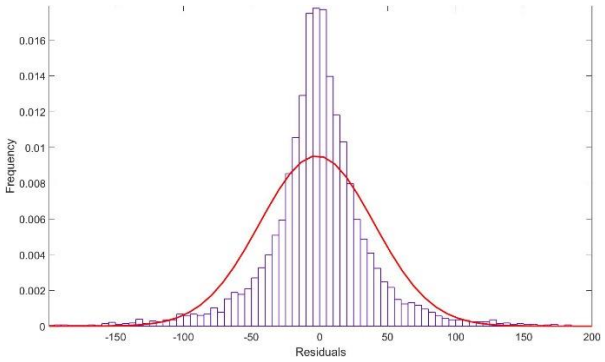


Figure 23 Residuals of GMCM and its gaussian fitting

It can be seen that the residuals remain highly concentrated around 0 and are even significantly lower than their fit to the Gaussian distribution.

4.3 Probabilistic assesment

As indicated in section 3.6, based on this GMCM, the p-value of the observed power for the observed value of wind speed at each instant of the unhealthy period is calculated. Fisher's test is performed by adjusting the number of independent tests to 3 in order to reduce the number of false positives. A significance level of 0.05 (corresponding to a 95% confidence bound) is used. The result of absolute yaw error and triggered alarm is shown in figure 24.

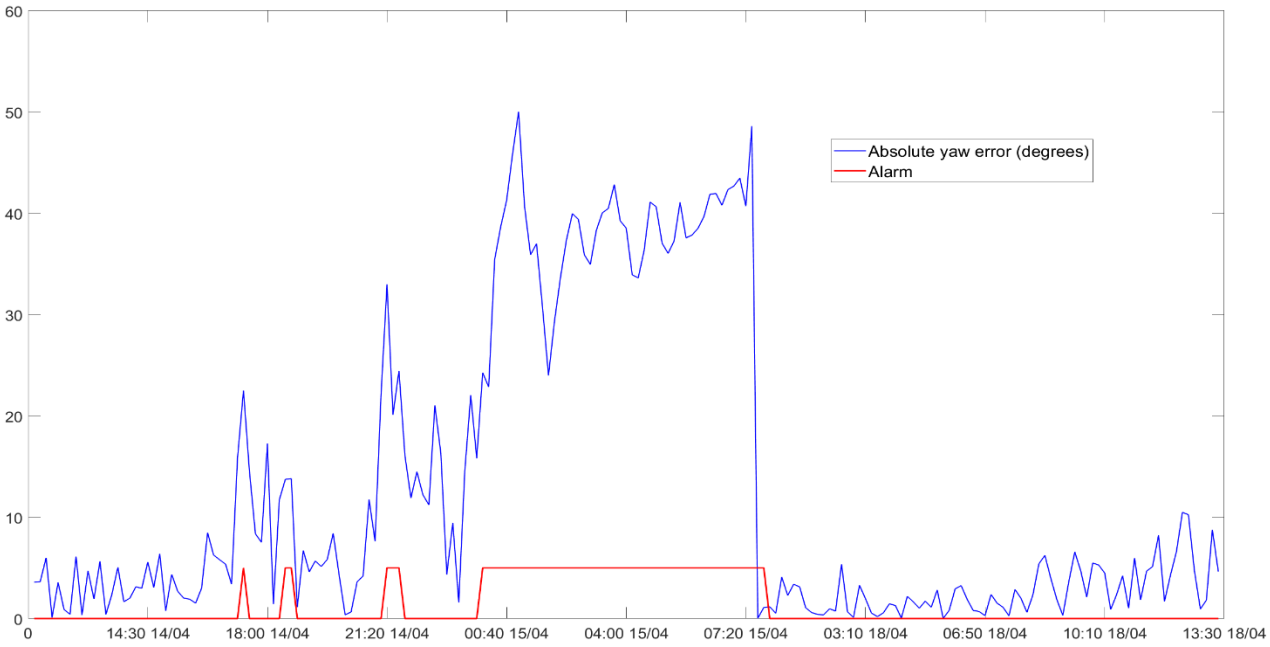


Figure 24 Yaw error detection with probabilistic approach of GMCM

Alarms were triggered at 17:20, 18:30-18-40, 21:20-21:40 and from 00:00 onwards, when the yaw error increased over a value and the described method detects suboptimal conditions in the wind turbine. In [25] , whose result is shown in figure 20, the failure of the yaw control had been determined at 17:00, so the first alarm is triggered 20 min after this, which is a rather early failure detection.

Yaw misalignment	Alarm triggered	First alarm since failure
From 17:00	17:20,18:30-18:40,21:20-21:40, from 00:00	20 min

Figure 25 Results for the failure detection

After each of the first three alarms, the alarm is reset as the wind direction changes and hence the yaw misalignment is reduced, which is to be expected as the proposed method only detects suboptimal turbine behavior. Later, at 00:00, when the yaw misalignment remains at high values, the alarm remains active. All this indicates the goodness of the obtained result.

It is worth noting that during the test period the number of false positives is kept at 0, which is another of the expected characteristics that any result that is developed must fulfil.

4.4 Comparison with other researchs

Now we can compare the result here obtained with others results found in the bibliography with the same dataset used in this project. In reference [25], it is used a probabilistic evaluation of a gaussian process (GP) model, the first alarm is triggered at 22:30. Also in [25], using GP but enhanced with the inclusion of the rotor speed, first triggered alarm occurs at 21:40. In reference [10], it is used the IEC binning method, the first alarm is triggered at 00:50. In reference [9], it is used probabilistic evaluation of a copulas method , the first alarm is triggered at 03:00

Model used	First alarm detected
GP model [25]	22:30
GP model with rotor speed [25]	21:40
IEC binning model [10]	00:50
Copulas model [9]	03:00
GMCM	17:20

Figure 26 Comparatives among models and assessment

In figures 27, 28 and 29 it can be seen the yaw error and results of the failure detection used in these quoted models, and the summary together with the results of this project in figure 26.

In figure 27 it can be seen that the proposed method of GP algorithm with inclusion of rotor speed is an improvement over the same method without inclusion of rotor speed, since on the one hand it decreases the detection time and on the other hand it eliminates the false positives that occurred in the latter. In both cases the detection of the fault only occurs when the yaw misalignment is already very evident and exceeds 20º.

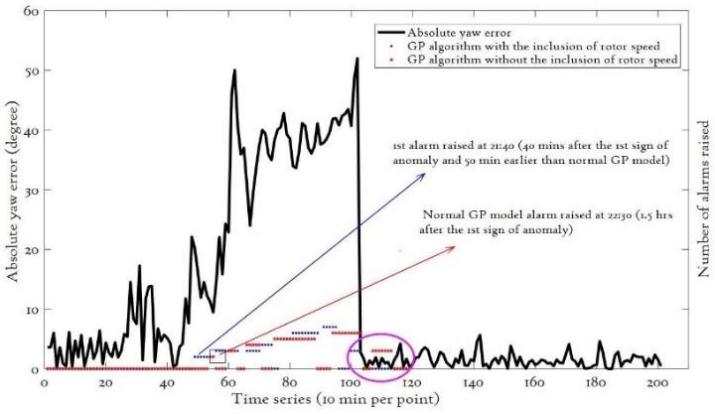


Figure 27 Yaw error detection with GP model [25]

Figure 28 shows how the proposed method of binned power curve takes longer to detect the yaw misalignment than the two previous methods. Regarding to the false positives, they are only detected two times and in localised points. Therefore, it is another valid method of anomaly detection, although it only acts when the failure is already evident.

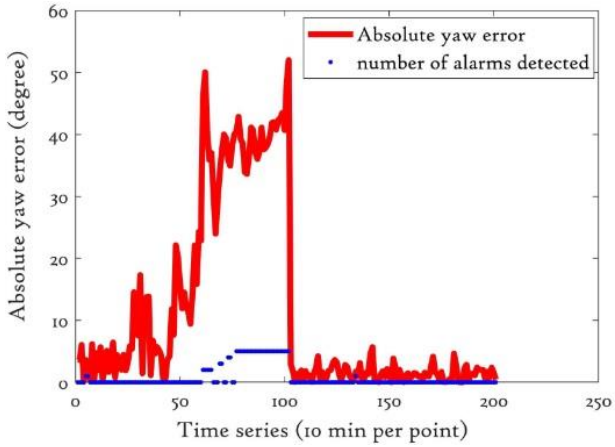


Figure 28 Yaw error detection using probabilistic assessment of binned power curve[10]

Figure 29 shows how the proposed method with copulas is also able to detect the anomaly but takes much longer to detect it. It also produces false positives, although these are located just after the elimination of the anomaly. Therefore, it can also be considered a valid method for the detection of anomalies, although not at an early stage.

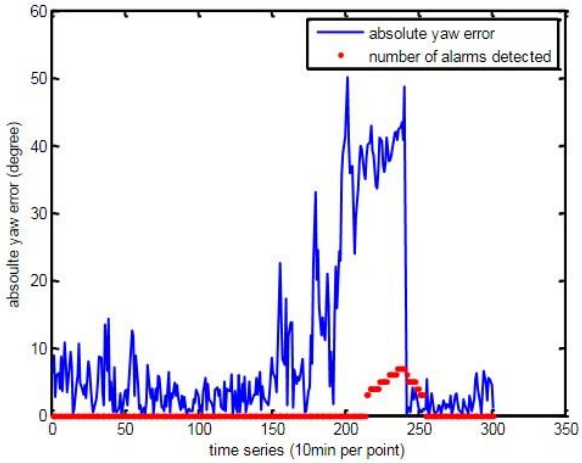


Figure 29 Yaw error detection with copula model[9]

From our results using the proposed model and method it can be concluded that it is also a valid method for anomaly detection. Moreover, it improves its efficiency with respect to the other

methods described in the literature, because it is able to detect yaw misalignment at earlier stages of the failure, even when this failure is not so evident or marked. In addition, the number of false positives is kept at 0 or close to 0 as with the other methods described.

CHAPTER 5: CONCLUSIONS AND FUTURE WORK

In the present project, the power curve has been presented as the most important indicator of the state of health of a wind turbine. It is in fact the best indicator of the performance of these systems, since a malfunction is immediately reflected in this curve. In addition to this, we can rely on SCADA acquisition systems that provide us with an automated system for monitoring conditions in real time. These two tools together constitute magnificent tools to be used in predictive maintenance procedures.

It has also been seen how GMCMs are a good mathematical tool for modelling the shape of the turbine power curve. Its main advantage is that it is data-driven, so it adapts faithfully to the specific conditions of each turbine and to the different operational regimes that make up the turbine, even when the dependence is non-linear.

The results show that the proposed probabilistic assessment together with the GMCM for this dataset gives a detection threshold for yaw misalignment of 20 min. This result is an improvement over other studies found in the literature. It should also be noted that, like the others considered, it does not produce false positives in this dataset. All this makes the present modelling and assessment method fast and reliable for a early failure detection.

Future work in line with the present project may include the following:

- Test the suitability of the GMCM with other types of failures that also involve a reduction in performance, such as: blade failures, pitch blade system faults, stiffness in mechanical parts, and so on.

- Check if it is possible to detect which system fails and activate the alarm.

- Include other SCADA variables as input parameters to the model to further refine the model, such as rotor speed, wind direction or atmospheric pressure.

List of figures

Figure 1 Off-shore wind farm [3] 1

Figure 2 Schematic of a wind turbine [26] 5

Figure 3 Power curve of wind turbine [4] 7

Figure 4 OEM power curve vs dataset 8

Figure 5 IEC standard power curve 9

Figure 6 Flowchart general methodology 13

Figure 7 Dataset used 14

Figure 8 Example of outliers in a power curve 16

Figure 9 Contour level of GMCM for outliers detection purpose 17

Figure 10 Power curve after outliers rejection 17

Figure 11 Samples of wind speed corrected 18

Figure 12 Elements of yaw control system [18] 19

Figure 13 Synthetic data for visualization of GMCM benefits [21] 21

Figure 14 GMM and GMCM fitting of figure 13 [21] 22

Figure 15 Comparison between types of copulas [23] 23

Figure 16 GMCM fit and CDF for Yalova example 25

Figure 17 Example of Z at wind speed of 4m/s 26

Figure 18 Absolute yaw error vs time 29

Figure 19 Unhealthy period 30

Figure 20 Wind & nacelle direction unhealthy period [25] 30

Figure 21 GMCM fit of healthy period 31

Figure 22 Dataset into copula space 31

Figure 23 Residuals of GMCM and its gaussian fitting 32

Figure 24 Yaw error detection with probabilistic approach of GMCM 33

Figure 25 Results for the failure detection 33

Figure 26 Comparatives among models and assessment 34

Figure 27 Yaw error detection with GP model [25] 34

Figure 28 Yaw error detection using probabilistic assessment of binned power curve[10] 35

Figure 29 Yaw error detection with copula model[9] 35

List of acronyms

O&M - Operation and Maintenance costs

SCADA - Supervisory Control And Data Acquisition

GMCM - Gaussian mixtures copula model

GMM - Gaussian mixtures model

OEM – Original Equipment Manufacturer

SOM – Self Organizing Map

PDF – Probability Density Function

CDF – Cumulative Density Function

WT – Wind Turbine

GP – Gaussian Process

WS- Wind speed

Bibliography and references

[1]*Strategic Energy Technology Information System* (SETIS) https://setis.ec.europa.eu/index_en

[2]*Energía eolica*. Instituto para la diversificación y ahorro energético https://www.idae.es/sites/default/files/documentos/publicaciones_idae/documentos_10374_energia_eolica_06_d9231f5c.pdf

[3] *Que es la energía eolica marina* – Iberdrola https://www.iberdrola.com/sostenibilidad/como-funciona-la-energia-eolica-marina

[4] Stephen, B.; Galloway, S. J.; McMillan, D.; Hill, D. C.; Infield, D. G.; *A Copula Model of Wind Turbine Performance*, IEEE Transactions on Power Systems 26 2, pp. 965–966, 2011.

[5] C. S. Gray, and S. J. Watson, *Physics of Failure approach to wind turbine condition based maintenance*, Wind Energy, 13: 395–405. doi: 10.1002/we.360, 2010.

[6] *Wind Turbines—Part 12-1: Wind energy generation systems - Part 12-1: Power performance measurements of electricity producing wind turbines*, IEC 61400-12-1:2017

[7] Simon Gill, Bruce Stephen, Stuart Galloway. *Wind Turbine Condition Assessment through Power Curve Copula Modeling*. IEEE Transactions on Sustainable Energy DOI: 10.1109/TSTE.2011.216716

[8] Bruce Stephen, Stuart J. Galloway. *A Copula Model of Wind Turbine Performance* DOI: 10.1109/TPWRS.2010.2073550

[9] Yue Wang, David G. Infield, Bruce Stephen, Stuart J. Galloway. *Power Curve Based Online Condition Monitoring for Wind Turbines*. DOI:10.13140/2.1.4492.9928

[10] Ravi Kumar Pandit, David Infield. *SCADA-based wind turbine anomaly detection using Gaussian process models for wind turbine condition monitoring purposes*. ISSN 1752-1416 (https://doi.org/10.1049/iet-rpg.2018.0156)

[11] Ravi Pandit and David Infield. *Gaussian Process Operational Curves for Wind Turbine Condition Monitoring*. DOI:10.3390/en11071631

[12] Simon Watson, Beth J. Xiang, Wenxian Yang, P.J. Tavner *Condition monitoring of the power output of wind turbine generators using wavelets*. DOI:10.1109/TEC.2010.2040083

[13] Zaher, A.; McArthur, S.; Infield, D.; Patel, Y. *Online wind turbine fault detection through automated SCADA data analysis*. Wind Energy 2009, 12, 574–593.

[14] Marvuglia, A.; Messineo, A. *Monitoring of wind farm's power curves using machine learning techniques.* Appl. Energy 2012, 98, 574–583

[15] Vidal, Y.; Pozo, F.; Tutivén, C. *Wind turbine multi-fault detection and classification based on SCADA data*. Energies 2018, 11, 3018

[16] Roger B. Nelsen. *An Introduction to Copulas.* Springer Book. ISBN: 978-0-387-28678-5

[17]Schlechtingen, M., Santos, I.F.: '*Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection*', Mech. Syst. Signal Process., 2011, 25, (5), pp. 1849–1875

[18] *Wind Energy* http://www.beaufortcourt.com/energy-generation/wind-energy

[19] *Power Loss from Yaw Misalignment* https://www.windesco.com/blog/yaw-misalignment-power-load

[20] Sklar, A., *Functions de repartition a n dimensions et leurs marges*. Publ. Inst. Statist. Univ. Paris, 1959. 8: pp. 229-231.

[21] A. Tewari, M. J. Giering and A. Raghunathan. *Parametric Characterization of Multimodal Distributions with Non-gaussian Modes*. 2011 IEEE 11th International Conference on Data Mining Workshops, 2011, pp. 286-292, doi: 10.1109/ICDMW.2011.135.

[22] Tewari Copula Mixture Model code [online]. Available: http://code.google.com/p/copula-mixture-model/downloads/list

[23] Wang, Y., Infield, D. G., Stephen, B. and Galloway, S. J. (2014), *Copula-based model for wind turbine power curve outlier rejection*, Wind Energ., 17, pages 1677– 1688, doi: 10.1002/we.1661

[24]Fisher, R.A.: '*Statistical methods for research workers*' (Oliver and Boyd, London, 1932)

[25] R. Pandit, D. Infield and T. Dodwell. *Operational Variables for Improving Industrial Wind Turbine Yaw Misalignment Early Fault Detection Capabilities Using Data-Driven Techniques*. IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-8, 2021, Art no. 2508108, doi: 10.1109/TIM.2021.3073698.

[26] Contreras Montoya, Hayyani, Issa, Ilinca, Ibrahim, Rezkallah. *Wind power plant planning and modeling,* Hybrid Renewable Energy Systems and Microgrids, Academic Press,2021,Pages 259-312, ISBN 9780128217245, https://doi.org/10.1016/B978-0-12-821724-5.00012-X.

[27] L. Zhang and Q. Yang, *A Method for Yaw Error Alignment of Wind Turbine Based on LiDAR*. *IEEE Access*, vol. 8, pp. 25052-25059, 2020, doi: 10.1109/ACCESS.2020.2969477.

[28] Zou M, Djokic SZ. A Review of Approaches for the Detection and Treatment of Outliers in Processing Wind Turbine and Wind Farm Measurements. *Energies*. 2020; 13(16):4228. https://doi.org/10.3390/en13164228

A. Appendix: code developed

Loading data

Import all the dataset

```
clear
dataset=readtable('2009Apr-Jun power curve.xls','Sheet',2);
```

Selection of rows I will need and convert to double type:

```
time=table2array(dataset(:,3));
time=datetime(time,'Format','yyyy-MM-dd HH:mm:00.000');
wind=table2array(dataset(:,4));
power=table2array(dataset(:,5));
yawerror=table2array(dataset(:,10));
temp=table2array(dataset(:,6));
```

Air density correction of wind speed

```
temp=273.15+double(temp);
density=1.225.*(288.15./temp);
Dens = [density./1.225].^(1/3);
wind=wind.*Dens;
```

Yaw Error transformation

Conversion of relative yaw error to absolute value and correction if it is more than ±180º

```
yawerror=abs(yawerror);
for i = 1:numel(yawerror)
  if yawerror(i) > 180
   yawerror(i) = 360 -yawerror(i);
  end
end
plot(time,yawerror)
dataset=timetable(time,wind,power,yawerror);
```

Zoom in to data with yaw error. It is stored in timetable “errordata”

```
TRnok = timerange('2009-04-14 11:20:00.000','2009-04-18 14:00:00.000');
errordata=dataset(TRnok,:);
plot(errordata.time,errordata.yawerror)
```

Healthy data

Dataset with healthy data is stored in timetable “data”:

```
TRok1 = timerange('2009-04-01 00:40:00.000','2009-04-14 11:10:00.000');
TRok2 = timerange('2009-04-18 14:10:00.000','2009-06-30 21:40:00.000');
okdata1=dataset(TRok1,:);
okdata2=dataset(TRok2,:);
data=[okdata1.power okdata1.wind; okdata2.power okdata2.wind];
plot(data(:,2),data(:,1),'.')
```

Plot bar:

Creation of the plot bar according to IEC

```
edges=1.75:0.5:12.75;
[Y,E] = discretize(data(:,2),edges);
media=splitapply(@mean,data(:,1),Y);
desv=splitapply(@std,data(:,1),Y);
edges(end)=[];
errorbar(edges+0.25,media,desv)
```

Reference model using GMCM

Based on code in reference [22] .Calculate the joint density probability function::

```
[N,d] = size(data); %N = number of datapoints, d = dimensions=2

% Defining the mixture model parameters
K = 2; % Number of modes
num_bins = 2^12; % needed for non-parametric marginal density estimation
num_GD_iterations = 200; % Max number of Gradient Descent iterations


% Obtaining the CDF values by fitting the marginal densities.
u = GMCM_marginals_learning(data,d,num_bins);

% Creating a Gaussian Copula Mixture (gmc) object
gmcObject = gmcdistribution([],[],[]);


% Estimating the parameters with the default options.
gmcObject_bestfit =
gmcObject.fit(u,K,d,N,'iteration',num2str(num_GD_iterations));


% scatter plot of the healthy dataset overlayed with the contours of fitted
% probability density values
[xgrid,ygrid,Z]=biVariateContourPlotsGMMCopula(data,gmcObject_bestfit,0,1024
,2,1);
%xgrid,ygrid is the grid with wind speed and power values respectively. Z
%contains the joint probability values evaluated in (xgrid,ygrid)

hold on
plot(data(:,2),data(:,1),'b.')
hold off
```

Probabilistic assessment

Code developed for assessment of the unhealthy data with yaw misalignment. Calculate p-value associated with every data test

```
%Creation of row with p-values
pvalue1=zeros(length(errordata.time),1);
errordata=addvars(errordata,pvalue1,'After','yawerror');

%First I search for the xgrid value (wind) closer to the point to test.
% Extract from the GMCM (using Z values calculated the PDF for the given
%xgrid(wind) in the test point.
```

```
%Calculate p-values of the power output of this test point. It is used a
$spline interpolation for the previous PDF
for i=1:length(errordata.time)
  [minimo,Columna]=min(abs(errordata.wind(i)-xgrid(1,:)));
  pol=spline(ygrid(:,Columna),Z(:,Columna));
  s=@(x)ppval(pol,x);
      errordata.pvalue1(i)=integral(s,min(min(ygrid)),errordata.power(i))/
      integral(s,min(min(ygrid)),max(max(ygrid)));
end
```

Now I calculate test Fisher of the unhealthy data using the previous pvalue1:

```
%Test Fisher stored in variable final_pvalue
final_pvalue=zeros(length(errordata.time),1);
errordata=addvars(errordata,final_pvalue,'After','pvalue1');
%Initialization of the firsts points
errordata.final_pvalue(1)=errordata.pvalue1(1);
errordata.final_pvalue(2)=errordata.pvalue1(2);

for i=3:length(errordata.time)
      pvals=[errordata.pvalue1(i) errordata.pvalue1(i-1)
      errordata.pvalue1(i-2)];
   chi_vals = -2.*log(pvals);
      errordata.final_pvalue(i) = 1- chi2cdf(sum(chi_vals),2*length(pvals));
end


%Alarm triggered if final_pvalue is less than significance level

alarm=zeros(length(errordata.time),1);
errordata=addvars(errordata,alarm,'After','final_pvalue');
for i=2:length(errordata.time)
    if errordata.final_pvalue(i)<0.05
      errordata.alarm(i)=5;
    end
end


%Plotting of the results

plot(errordata.time,errordata.yawerror,'b',errordata.time,errordata.alarm,'r
-')
```