Wind turbine generator prognostics using field SCADA data

Peter, R.A.; Zappalá, D.; Schamboeck, Verena; Watson, S.J.

DOI
10.1088/1742-6596/2265/3/032111

Publication date
2022

Document Version
Final published version

Published in
Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Wind turbine generator prognostics using field SCADA data

To cite this article: Rudolph Peter *et al* 2022 *J. Phys.: Conf. Ser.* **2265** 032111

Journal of Physics: Conference Series **2265** (2022) 032111 doi:10.1088/1742-6596/2265/3/032111

Wind turbine generator prognostics using field SCADA data

Rudolph Peter[1], **Donatella Zappalá**[1], **Verena Schamboeck**[2], **Simon J Watson**[1]

[1] TU Delft, Kluyverweg 1, 2629 HS Delft, Netherlands
[2] Vattenfall, Hoekenrode 8, 1102 BR Amsterdam, Netherlands

E-mail: `rudolphpeter1997@gmail.com`, `d.zappala@tudelft.nl`, `verena.schamboeck@gmx.at`, `s.j.watson@tudelft.nl`

Abstract. This paper presents a novel prognostic method to estimate the remaining useful life (RUL) of generators using the SCADA (Supervisory Control And Data Acquisition) systems installed in wind turbines. A data-driven wind turbine anomaly classification method is developed. The anomalies are quantified into a health indicator to measure the component degradation over time. An Autoregressive Integrated Moving Average (ARIMA) time series forecasting technique is then applied to predict the RUL of the wind turbine generator. The proposed method has been validated using industry field data showing accurate predictions of RUL with a 21 day lead time for maintenance of the turbine.

1. Introduction

Technological developments in wind energy have made wind one of the cheapest and most sustainable sources of energy. The rapid growth of wind energy capacity, especially offshore increases the challenges in conducting operations and maintenance (O&M) of the turbines [1]. According to Stehly et al. [2] O&M costs account for up to 35% of the total cost of offshore wind energy. Given the consistent increase in the installed capacity worldwide, all the operational challenges and the costs associated with the O&M of wind turbines, current research efforts are aimed at reducing downtime and increasing the turbine reliability, availability and performance. Accurate prognostics of failures of wind turbine components is the key to assuring reliable and efficient wind farm performance and it can contribute to significantly reducing O&M costs by providing sufficient lead time to procure spare components and create maintenance plans. Although advances have been made in fault diagnostics, very little research using real operational data has been carried out in the field of wind turbine fault prognostics [3]. Current diagnostic and prognostic solutions for wind turbines rely on the use of purpose-built, expensive condition monitoring systems (CMS). Supervisory Control And Data Acquisition (SCADA) systems are standard installations on turbines that provide a range of low frequency (typically 10-min averaged values) measurements to monitor the turbine operation and performance. SCADA is a valuable low-cost monitoring system that can significantly contribute to O&M cost reduction. However, despite their potential, SCADA data are currently not widely used for wind turbine failure prognosis. Current research on wind turbine prognostics is heavily focused on bearings and gearbox, mainly using CMS data [4]. The generator is one of the most expensive components of the turbine and it fails just as often as the gearbox, with failures causing long downtimes. Therefore, it is an important component to perform prognostics on [5].

The objective of this work is to develop a novel wind turbine generator prognostic framework using field SCADA data. The data for this research are supplied by Vattenfall, a utility company that owns and operates wind farms in Northern Europe. Through the use of targeted SCADA signals and machine learning techniques, the failure behaviour is quantified into a health indicator which can be forecasted to estimate the generator remaining useful life (RUL). The developed prognostic model is then validated on several generator failures using field data to understand its advantages and limitations.

2. Wind Turbine Prognostics

Rezamand et al. [4] categorizes prognostic methods into four different classes, physics-based, AI-based, stochastic-based and hybrid prognostic techniques. Physics-based prognostic methods build mathematical models to estimate the RUL of components. They can potentially be very accurate but they usually require detailed information about the components of the turbine which is not usually easily available [4]. AI-based prognostics forecast the failure of a turbine component by employing machine learning (ML) techniques to model the failure of components from the observational data. They are excellent at modeling nonlinear complex systems but require a large volume of training data that is representative of all different operating conditions [6]. Stochastic-based prognostics employ probability theory as the basis of their methods. They are capable of providing long-term RUL forecasts along with the uncertainty in the forecast however, they require a statistically significant sample size for each failure mode of a component to make a reliable prediction [6]. Hybrid prognostics methods are those methods that are constructed using a combination of the previously described methods. These methods are being studied extensively because by combining more than one technique the drawbacks of individual techniques can be overcome [4]. The prognostic framework proposed in this work is a hybrid approach that combines AI-based and stochastic prognostics.

Zhao et al. [7] proposed a hybrid method to forecast the failure of wind turbine generators using SCADA data. By using a clustering algorithm called DBSCAN (Density-based spatial clustering of applications with noise), the authors detect anomalies in the data and define a health indicator, which is then forecasted using an ARIMA model. However, the DBSCAN algorithm is not deterministic and does not work well with clusters of varying density. Since the effectiveness of DBSCAN-based clustering is highly dependent on the dataset, the method proposed by Zhao et al. [7] does not generalize well. To overcome these limitations and to build a more robust prognostic framework, the approach proposed in this paper replaces the DBSCAN with a classification-based anomaly detection algorithm.

3. Methodology

The goal of the prognostic model in this work is to forecast the failure of the generator. To do this, first, the health of the generator needs to be quantified. This is done by introducing a health indicator called Anomaly Operation Index (AOI). The AOI is given by the ratio between the number of faulty data points and the total number of data points within a defined time window. Distinguishing the normal instances in the data from the faulty instances (anomalies) can be achieved by using classification ML algorithms. The trend of the AOI can then be forecast to predict its status in the future. Based on the AOI forecasts, the operator can decide when to stop the turbine and perform maintenance. The prognostic approach is divided into three major layers and its schematic is shown in Figure 1.

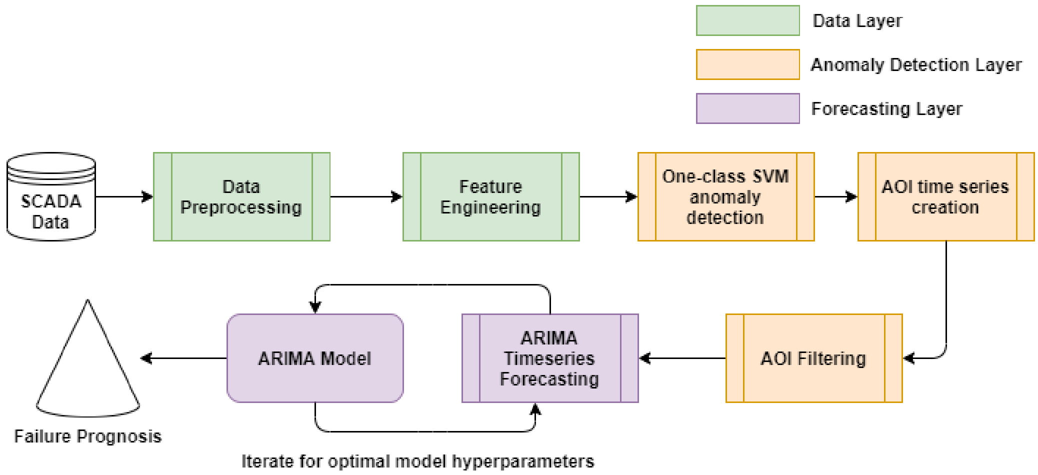

Figure 1. Prognostic Model Architecture.

3.1. Data Layer

In this layer, the raw SCADA data is pre-processed to be effectively used by the ML algorithms in the following layers. The SCADA system collects numerous signals from the turbine, however not all the signals are relevant to the generator's health. Therefore, the key signals selected as representative of the health and performance of the generator are: the drive end bearing temperature, the non-drive end bearing temperature, the cooling water temperature, the rotor temperature, and the slip ring temperature. In addition, the active power, the generator rotational speed, and the nacelle temperature have been selected to quantify the turbine's operational performance. Finally, the wind speed and the ambient temperature have been selected to represent the external conditions.

3.1.1. Data Preprocessing. SCADA data can contain errors caused by malfunctions in the data collection system and by sensor faults. The errors may be in the form of data gaps, out of range/missing values, etc. The preprocessing stage involves data exploration and filtering out errors, using criteria based on the operator's experience. Instances of turbine stoppages and curtailments are also filtered out in this step as the classification algorithm may classify them as anomalies.

3.1.2. Feature Engineering. Following the selection of signals from the SCADA data, feature engineering is performed on the data. This is the process of transforming raw data into features that enhance the performance of ML algorithms, which are highly dependent on data quality. The two major improvements that can be achieved through feature engineering are better model predictability and reduced complexity, which result in faster run times. In this work, the feature engineering performed includes the normalisation of the temperature signals to the nacelle temperature, the selection of the training and testing datasets, and the data dimensionality reduction.

Temperature Normalisation: Generator temperature signals are sensitive to the ambient temperature as well as the turbine conditions. To negate the effect of the nacelle temperature, the generator temperature measurements are normalised with respect to the turbine nacelle temperature by calculating the difference between the two signals.

Training and testing data selection: The training data for this model is selected from a healthy turbine (without generator faults) located next to the failed turbine in the same wind farm. The training turbine is of the same model as the faulty turbine. It is selected as a neighbouring turbine as it is likely to experience similar operating conditions (ambient temperature, wind speed, etc.) as the failed turbine. Ideally, the training dataset should be as large as possible to prevent overfitting of the data and to better generalise the ML algorithm. In this work, the training data is selected for a period of 1 year, prior to the date of failure of the faulty turbine. A period of 1 year is the minimum period required to capture all the seasonal effects throughout the year. A future consideration could be to use a larger training set that is built on data from several turbines. This would reduce the risk of the model learning turbine-specific behaviours that do not occur on the faulty turbine.

The testing data for this model is selected from the turbine with the generator failure for a period of 4 years prior to the failure. A longer testing period is selected to identify when the fault begins to develop and to check whether there may be any seasonal trends that appear in the health indicator.

Dimensionality Reduction: The transformation of high-dimensional data into a meaningful representation of reduced dimensionality is called dimensionality reduction (DR) [8]. DR reduces the number of features and hence the complexity of the model, decreasing the chances of overfitting and poor model performance. DR provides other benefits such as faster training times, lower storage requirements, and noise reduction.

The DR method used in this research is based on the independent component analysis (ICA) using the FastICA algorithm developed by Hyvarinen and Oja [8]. As the ICA is not the major contribution of this work, the details of its application can be referred to in [8].

3.2. Anomaly Detection Layer

In this layer, the optimised SCADA data set is fed into a one-class Support Vector Machine (SVM) algorithm to classify normal data and anomalies. The classified data is then used to calculate the AOI and quantify the health of the generator as a continuous time series.

3.2.1. One-class SVM Anomaly Detection. Several clustering/classification algorithms can be adapted to perform the anomaly detection. Classification is chosen over clustering in this research as there is no prior knowledge on whether the faulty data points are strictly separated into a cluster or the number of clusters to expect. With classification however, the algorithm can be trained on data that is known to be representative of normal operation and therefore the anomalous behaviour of the testing data can be correctly categorised by the algorithm. One-class SVM is an implementation of SVM that is dedicated to anomaly detection and it has been selected due to its effectiveness in classifying high dimensional data [9]. The one-class SVM algorithm is trained with the training data set from the neighbouring healthy turbine. The historical data from the failed turbine is then applied to the trained one-class SVM algorithm to classify the data into a normal set and an anomaly set.

3.2.2. Anomaly Operation Index Calculation. The normal set represents the normal operation of the turbine. The anomaly set represents the turbine's faulty operation. To quantify the faulty operation of the turbine with respect to a particular time window, the AOI is defined as the ratio of anomaly data points to total data points as:

$$AOI(t) = \frac{N_{anomaly}(t)}{N_{total}(t)} \tag{1}$$

where $N_{anomaly}(t)$ and $N_{total}(t)$ are the total number of anomalies and the total number of points, respectively, in the time window t.

SCADA data is provided in 10-minute intervals and the data points obtained after the preprocessing phase are classified as normal or anomalous by the one-class SVM algorithm.

Based on practical considerations, in this work a 1-day time window is chosen to calculate the AOI. The maintenance of wind farms requires several days of planning, especially offshore, to source components and arrange transportation and labour. Therefore, a useful unit to present the component RUL is in days. By using a 1-day aggregation window a daily AOI time series is created.

The higher is the AOI value, the higher is the proportion of faulty to healthy operation points which is representative of worsening generator performance. A continuous AOI time series can then be used to describe the generator performance trajectory. Assuming that the generator failure is an accumulated process, it is expected that as the turbine fault progresses towards a failure larger values of AOI are observed.

3.2.3. AOI time series filtering. Due to the stochastic nature of the failure process, the AOI time series will exhibit fluctuations. They may also contain time gaps due to the filtering of the raw data. Since forecasting techniques achieve better results when there is not much noise in the data and the curve is smooth, a simple moving average (SMA) is applied to the AOI time series so that the overall trend in the data can be captured while removing the outliers and noise. The averaging period is optimised by looping through different averaging windows and minimising the mean squared error between the forecasted and actual values.

3.3. Forecasting Layer

In this layer, the AOI time series created in the anomaly detection layer is filtered and forecasted using an ARIMA model. The forecast is used to estimate the RUL of the generator.

3.3.1. ARIMA time series forecasting. In order to forecast the AOI an ARIMA time series prediction technique is adopted. The theory behind this forecasting technique can be read in Box et al.'s work [10]. The two major categories of time series prediction methods are statistical and ML-based forecasting. The selection of the most appropriate forecasting technique depends on the type of dataset at hand. Makridakis et al. [11] research compared the performance of eight ML-based forecasting methods with that of eight statistical methods. Their results show that the ARIMA method outperforms the ML methods. They also note that statistical methods provide better interpretability of results as the forecasts are provided with uncertainty. Cerqueira et al. [12] argues that the research presented in [11] is only valid for datasets of small sizes while with larger datasets ML methods show improved performance. Nevertheless, in the case of multi-step forecasting, the ARIMA method shows better performance [12]. The research presented in this paper requires the multi-step ahead forecasting of a univariate time series. Since it is not evident from the literature that a ML-based method will lead to better forecasting performance, in order to preserve the interpretability feature of statistical approaches the ARIMA method is chosen. The implementation of ARIMA is performed in Python using the auto_arima function in the Pmdarima package. The auto_arima function performs a hyperparameter optimisation to find the best fit of the ARIMA model which is then used for the AOI forecast.

3.3.2. RUL Estimation Once the optimised ARIMA model is created, the generator RUL can be estimated. This is done by applying an AOI failure threshold value. In the case studies presented in this research, the threshold value is the value of the filtered AOI curve at the point of failure as this is the critical AOI value. Several points in time prior to the turbine's failure (7,14,21,30,60 days) are selected and forecasts are produced. The multi-day forecasts are produced through recursion i.e the ARIMA model produces a prediction for one-step and subsequent predictions are used as inputs into the same model to predict the following time step. The forecast is then checked against the AOI threshold value. When the forecast crosses the threshold value, that time step is determined to be the estimated failure point of the turbine. The number of days between the start of the forecast and the failure point represents the estimated RUL of the turbine.

For practical use, when neither the actual failure date of the turbine nor the failure AOI is known, the threshold can be determined based on historical AOI values for several similar failures that have occurred in the past. This threshold value can be adjusted according to the risk level that is conducive to the operator. If the operator requires a conservative RUL estimate, a lower threshold can be set and vice versa.

4. Results

4.1. Dataset Description

The SCADA dataset used in this work comes from an offshore wind farm located in the North Sea that has approximately 100 wind turbines. The wind turbines are geared and have a rated power of 3 MW each. The majority of the SCADA signals are stored as 10-min averages while only a few signals are available at 1 Hz frequency. As the majority of features selected in this work are temperature signals which have a large inertia, their low 10-minute sampling frequency is not expected to impact the prognosis forecasting.

The prognostic framework developed in this work is first applied to data related to a major failure of the generator non-drive end (NDE) bearing that occurred on turbine 2 in February 2019. The root cause analysis conducted found that the failure was caused by an electrical fault in the generator windings which led to high NDE bearing temperatures.

4.2. Data Layer Output

The training dataset used in the analysis is from a neighbouring turbine (Turbine 1) which did not experience failures in the period February 2018 - February 2019. The testing dataset is from the failed turbine (Turbine 2) and refers to a period of four years prior to its failure (February 2015 - February 2019). From now on the training Turbine 1 is referred to as the 'healthy turbine' and the test Turbine 2 as the 'faulty turbine'. The original total number of data points in the combined datasets is 262,800. Around 18% of them are filtered out during the pre-processing phase.

After applying the FastICA algorithm, the data set initially containing 10 features is transformed into five independent components. 2-D scatter plots of the independent components, as shown in Figures 2 & 3 for the case of components 3 and 5, are used to visualise the behaviour of the healthy and the faulty turbines.

The figures show a clear difference in behaviour between the two turbines. The healthy turbine exhibits a smaller area of operation while the faulty turbine shows data that is more widely distributed. Both turbines have an overlapping area which can be considered as the region of healthy operation for both turbines.

4.3. Anomaly Detection Layer Output

The one-class SVM algorithm is expected to form a classification boundary around the green region shown in Figure 2 and classify all the points outside that area as outliers. This is confirmed by Figure 4 showing the results of the one-class SVM classification for the failed turbine.

The top plot in Figure 5 shows the calculated daily AOI while the bottom plot shows the averaged AOI values, calculated by filtering the daily AOI curve with a 90 day SMA, as described in Section 3.2.3. While the raw AOI behaviour is very noisy and is difficult to visualise a trend toward failure, the SMA curve shows increasing values when approaching the point of failure. From December 2016, the averaged AOI shows a seasonal trend with peaks in the winter and troughs in the summer. The trough in the summer of 2017 is deeper than in 2018, indicating that the problem may have worsened in 2018. This seasonal effect of the AOI can be explained

Journal of Physics: Conference Series **2265** (2022) 032111 doi:10.1088/1742-6596/2265/3/032111

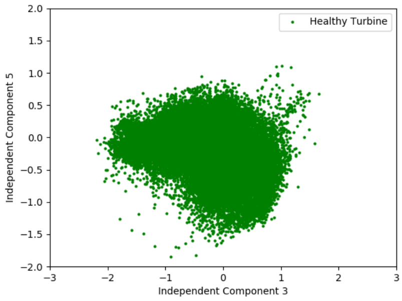

Figure 2. FastICA components 3 and 5 for Turbine 1.

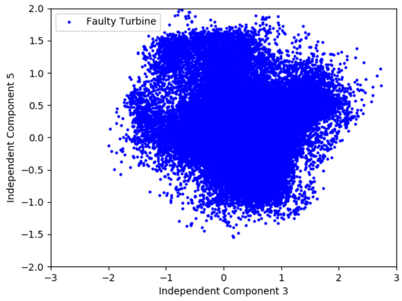

Figure 3. FastICA components 3 and 5 for Turbine 2.

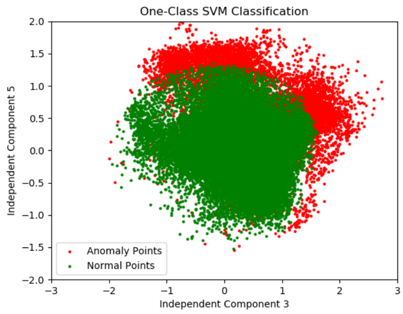

Figure 4. Anomaly classification for Turbine 2.

by the fact that the anomalies are captured by the model during periods of high wind speeds, which occur more often in winter. During high wind speeds, the turbines produce more power and therefore the temperatures in the generator are higher.

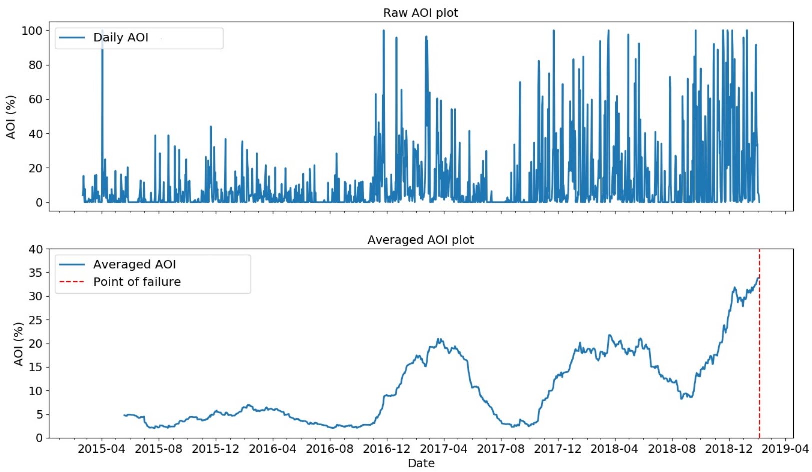

Figure 5. Daily (top) and averaged (bottom) AOI for Turbine 2.

4.4. Forecasting Layer Output

The RUL prediction accuracy for a certain time step t is calculated as:

$$RUL\ Accuracy(\%) = 1 - \frac{|T_t^* - T_t|}{T_t} \tag{2}$$

where T_t^* is the estimated RUL and T_t is the actual RUL. As described in Section 3.3.2, The ARIMA AOI forecasts are done 7, 14, 21, 30 and 60 days prior to the failure of the generator. The parameters (p,d,q) of the ARIMA model are (1,1,3). The 21-day forecast plotted in Figure 6 shows that the failure date is predicted just one day after the actual failure date. The closer to the failure the forecast is done the more accurate it is, as shown in Figure 7. For practical use, different warning levels with increasing severity based on the duration of the RUL can be sent to the operator.

Figure 8 shows the results of the sensitivity analysis performed to optimise the AOI SMA averaging window. It is clear that the larger the averaging window is the lower is the forecast error, expressed in terms of both mean squared error (MSE) and the symmetric mean absolute percentage error (sMAPE). This is because a larger window reduces the intrinsic noise and hence, the ARIMA model is able to pick up the trend in the data more clearly. However, this does not mean that the averaging window size can be increased indefinitely. The drawback of a large averaging window is that a larger number of past values influences the averaged AOI. Therefore, as the turbine progresses to failure, more gradual increasing slopes will be observed. This is problematic because mechanical faults in the generator tend to progress to failure quickly. By using a larger filtering window a steep rise in AOI can not be captured, leading to an overestimation of the turbine RUL.

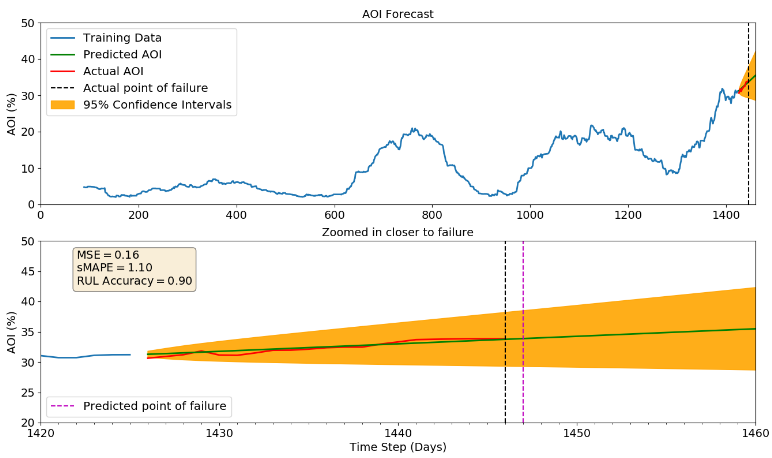

Figure 6. ARIMA 21 day forecast of Turbine 2.

4.5. Validation

To investigate the ability of the proposed model in forecasting different types of failures, two other cases studies were analysed. They refer to generator failures that occurred in Turbine 3 and 4 of the same wind farm.

In August 2020, the generator of Turbine 3 failed due to high temperatures in the generator. Figure 9 shows the turbine 21-day AOI forecast. The failure behaviour is different from the one

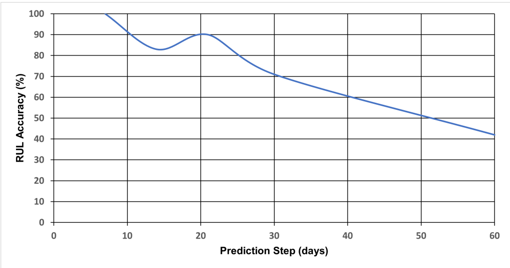

Figure 7. RUL accuracy vs. prediction time.

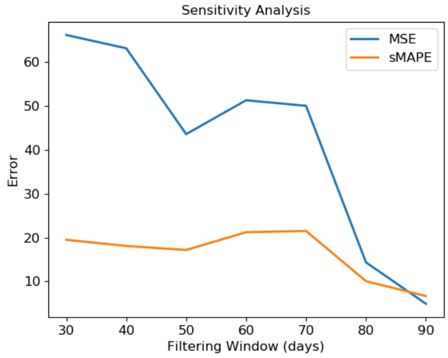

Figure 8. Averaging window sensitivity analysis.

observed for Turbine 2. For this turbine the AOI curve has a steeper slope, indicating that the problem developed more rapidly, and it does not show any seasonal trend. The model forecasts the failure of the turbine three days after the actual failure date with a forecast accuracy of 86%.

The generator failure of Turbine 4 occurred in February 2020. The root cause of this failure is unknown. Figure 10 shows the behaviour of the averaged AOI. An increasing trend towards failure is not visible, hence the model is unable to forecast the failure. However, an abnormally steep AOI increase can be observed seven months prior to the failure, well before the SCADA alarm for that fault was triggered. This increase corresponds to the irregular behaviour of the slip ring temperature, which was the main feature causing the anomalies. The AOI values decrease only after the failure. This shows that failures that occur suddenly without a prior degradation phase, as in this case, cannot be predicted by the ARIMA. However, the proposed model can still warn the operator of irregular turbine behaviour. This demonstrates the utility of the proposed framework to perform both fault detection and prognostics. In practice, this model can also be used for fault detection by setting appropriate AOI thresholds. When these values are exceeded a warning can be sent to the operator to trigger further investigation and appropriate maintenance actions.

It can be noted here that the three different turbines have slightly different AOI values at failure and this is because each turbine exhibits a different failure mode. In general, when more data is available, a failure threshold AOI value could be estimated by training the model on multiple failures and calculating an average AOI value. When the AOI ARIMA forecast exceeds this threshold the turbine can be considered to have failed.

5. Conclusions

This paper presents a novel wind turbine generator prognostic framework using field SCADA data. Through the use of targeted SCADA signals and the one-class SVM algorithm, the normal and faulty behaviour of the turbine's generator is classified. The faulty behaviour is quantified by using the AOI health indicator which is then forecasted using an ARIMA technique to estimate the generator RUL. The developed prognostic model is validated on three generator failures to understand its advantages and limitations. The results show that the model successfully forecasts failures of the generator while providing a 21-day lead time to the operators to plan the necessary maintenance action. The lead time necessary to carry out maintenance of wind turbines s subjective to the type of failure, location of turbine, component availability, etc. It also depends on the failure mode and therefore a direct comparison cannot be made with

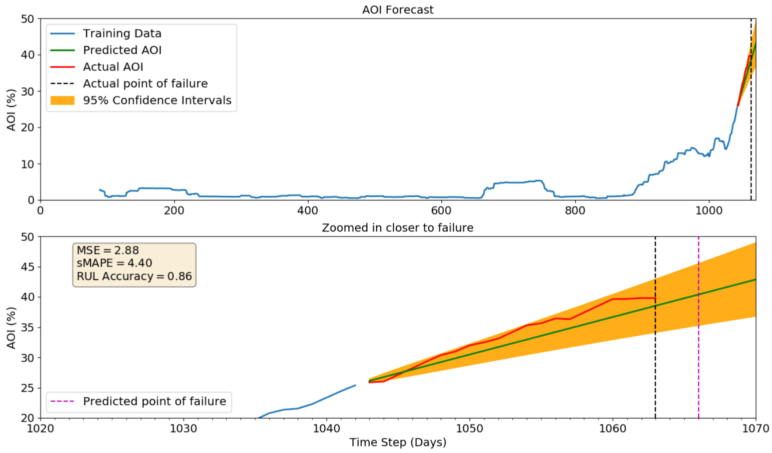

Figure 9. ARIMA 21-day forecast of the Turbine 3.

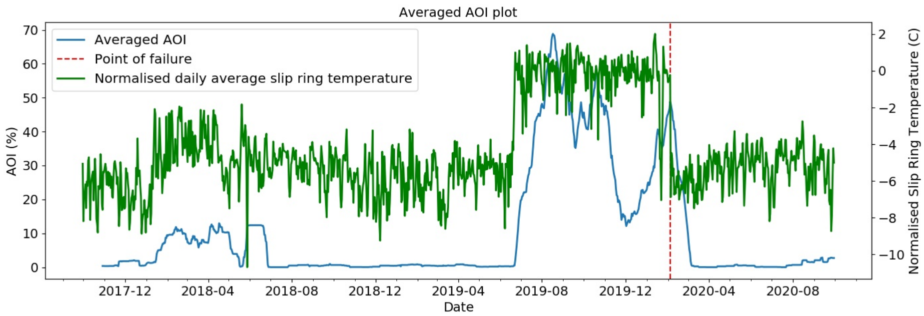

Figure 10. AOI and generator slip ring temperature of Turbine 4.

other prognostic methods. The proposed prognostic approach can help prevent catastrophic failures and provide advanced warning of faults which can contribute to reducing the O&M costs. This methodology is not failure specific and can be used to forecast a variety of failure modes, depending on the SCADA signal availability. Therefore, if more SCADA signals relating to the health of the generator are available more failure modes may be forecasted. When the model is unsuccessful at forecasting the failure, it can still detect faulty behaviour and provide timely warnings to the operators. The prognostic framework is versatile as it can be tailored to different drive train turbine components if relevant SCADA signals are available. This approach can be easily implemented in the field as a fault detection and prognostic tool as it relies only on the use of signals from standard widely adopted SCADA systems without requiring any expensive additional hardware.

References

[1] Global Wind Energy Council 2021 *GWEC: Global Wind Report 2021*
[2] Stehly T and others 2020 *2019 Cost of Wind Energy Review* (Golden,CO: National Renewable Energy Lab.(NREL))
[3] Leite G, Araujo A, Rosas P 2018 *Renewable and Sustainable Energy Reviews* **81** 1917-25
[4] Rezamand M, Kordestani M, Carriveau R, Ting D, Orchard M, Saif M 2020 *IEEE Transactions on Instrumentation and Measurement* **69** 9306-28
[5] Yang W, Tavner P, Crabtree C, Feng Y,Qiu Y 2014 *Wind Energ.* **17** 673-693

[6] Sikorska J, Hodkiewicz M, Ma L 2011 *Mechanical systems and signal processing* **25** 1803-36
[7] Zhao Y and others 2016 *Fault Prognosis of Wind Turbine Generator Using SCADA Data* (Denver,CO: North American Power Symposium (NAPS)) 1-6
[8] Hyvarinen A, Oja E 2000 *Neural networks* **13** 411-430
[9] Sutharan S 2016 *Support Vector Machine* (Machine learning models and algorithms for big data classification) (Boston,MA: Springer) pp 207-235
[10] Box G and others 2015 *Time series analysis: forecasting and control* (Hoboken,NJ: Wiley)
[11] Makridakis S, Spiliotis E, Assimakopoulos V 2018 *PloS one* **13** e0194889
[12] Cerqueira V, Torgo L, Soares C 2019 *arXiv preprint arXiv:1909.13316*