Available online at www.sciencedirect.com

ScienceDirect

Energy Procedia 137 (2017) 571–578

Energy Procedia

www.elsevier.com/locate/procedia

14th Deep Sea Offshore Wind R&D Conference, EERA DeepWind'2017, 18-20 January 2017, Trondheim, Norway

Optimisation of Data Acquisition in Wind Turbines with Data-Driven Conversion Functions for Sensor Measurements

L. Colone [*,a], M. Reder[*,b], J. Tautz-Weinert[*,c], J.J. Melero[b], A. Natarajan[a], S.J. Watson[c]

[*] *Shared first authorship - authors contributed equally to the publication and are presented in alphabetical order.*

[a] *Technical University of Denmark, Frederiksborgvej 4000, Roskilde, Denmark*
[b] *CIRCE - Universidad de Zaragoza, C/ Mariano Esquillor 15, 50018, Zaragoza, Spain*
[c] *CREST - Loughborough University, Holywell Park, Loughborough, LE113TU, UK*

Abstract

Operation and Maintenance (O&M) is an important cost driver of modern wind turbines. Condition monitoring (CM) allows the implementation of predictive O&M strategies helping to reduce costs. In this work a novel approach for wind turbine condition monitoring is proposed focusing on synergistic effects of coexisting sensing technologies. The main objective is to understand the predictability of signals using information from other measurements recorded at different locations of the turbine. The approach is based on a multi-step procedure to pre-process data, train a set of conversion functions and evaluate their performance. A subsequent sensitivity analysis measuring the impact of the input variables on the predicted response reveals hidden relationships between signals. The concept feasibility is tested in a case study using Supervisory Control And Data Acquisition (SCADA) data from an offshore turbine.

Peer-review under responsibility of SINTEF Energi AS.

Keywords: Wind Turbine, Condition Monitoring, SCADA, Data Optimisation, Data Mining, Operation and Maintenance (O&M);

1. Introduction

The costs associated with Operation and Maintenance (O&M) of wind turbines account for about 10-15% of the overall energy generation cost for onshore [1] and 25-30% for offshore wind turbines [2]. For wind farms approaching the end of life the O&M costs may rise up to 35% [3]. Currently, the wind industry is incurring significant numbers of main component failures, causing large downtimes and consequently loss of power production [4], [5], [6]. Throughout many years of experience in other industrial sectors, condition based maintenance (CBM) has become an established and cost-effective maintenance strategy [7], [8]. The need for cost-effective maintenance is further intensified

[*] Corresponding authors
E-mail address: lcol@dtu.dk, mreder@fcirce.es, j.tautz-weinert@lboro.ac.uk

1876-6102 © 2017 The Authors. Published by Elsevier Ltd.
Peer-review under responsibility of SINTEF Energi AS.
10.1016/j.egypro.2017.10.386

by the increasing installation of wind turbines offshore, where logistics are more difficult. This implies an increased reliance on remote sensing systems for health assessment in the transition from corrective to predictive maintenance.

Early condition monitoring of wind turbines was based on the high frequency signals provided by a dedicated Condition Monitoring System (CMS), in particular acceleration measurements (cf. [7]). The advantages and economic benefits of CMS in the wind energy sector were analysed for instance in [9]-[10]. Current trends also provide clear evidence of an increasing exploitation of Supervisory Control And Data Acquisition (SCADA) data for monitoring purposes, thanks to its economic advantages and well established online data management. Following the latest development in the field with regards to wind energy applications, artificial intelligence systems are receiving greater consideration. An example of application is presented by Kusiak and Verma where genetic programming is employed to predict blade pitch faults as early as an hour before occurrence [11]. Bangalore and Tjernberg detected gearbox bearing damage with artificial neural networks (ANNs) one week earlier than the CMS [12]. A comprehensive review of the main advances in this area is provided in [13].

There is a potential to widen the monitoring concept by taking advantage from combined information of operational and condition monitoring data. However, wind farm operators and manufacturers claim to gather extensive data from wind farms while lacking the capability to translate data into useful information for decision making. This motivates further research for optimisation and standardisation of monitoring systems [14]. In this context, less effort has been made to investigate synergistic effects of coexisting sensing technologies, e.g. SCADA and CMS. Hence, modern approaches should focus on analysing the correlation between signals, in the attempt to enable better understanding of the measurement data and eventually exclude irrelevant input variables.

The initial idea of the concept presented in this paper was developed during the 1st *Joint Industrial Workshop* (JIW) within the European Union's H2020 project AWESOME, as documented in [15]. This work is a continuation and extension of the main idea presented in the workshop. A case study is carried out testing different techniques on field data. The primary objective is the optimisation of the data acquisition by taking advantage of correlated signals and mining algorithms. The base of the methodology is the prediction of certain variables using a set of conversion functions between measurements. The approach is fully data driven, which implies that it is not relying on physical models. The next section of this paper includes an outline of the proposed approach, followed by details of a case study and results. The last section provides an outlook of ongoing and recommended future research in this field.

2. Methodology

A framework is proposed to investigate relationships between coexisting measurements to reveal potentially helpful correlations and synergistic effects. This hidden information will be identified by the evaluation of data-driven conversion functions. To allow all possible interactions, selecting the input-output relations is not limited to a physical understanding of the system. On the contrary, each available signal $x_1, x_2, x_3 ... x_{n-1}$ has to be used as an input for modelling one of the other signals (x_i). Only the target signal itself is excluded from the input set, in order to discard trivial conversion functions. Each signal acts once as the target, resulting in n multiple input and single output conversion functions predicting with an error ϵ.

$$x_i = f_i(x \in X \setminus x_i) + \epsilon_i \qquad \text{with } X = \{x_1, x_2, x_3, ..., x_n\} \text{ and } i = 1, 2, 3, ..., n \tag{1}$$

Figure 1 illustrates a single exemplary conversion function in a wind turbine drive train. The investigation of synergistic effects is based on three main steps:

1. Pre-process and extract features of training data,
2. Build n conversion functions,
3. Evaluate conversion functions.

In the first step, measurement data are prepared to be used to set up and feed the conversion functions. Pre-processing has to include checking for missing signals and invalid values. Duplicated or non-working sensors are excluded. As SCADA signals, which are commonly provided in 10 minute resolution, will be combined with dedicated CMS measurements at high sampling rate, a common working frequency has to be determined. It might be

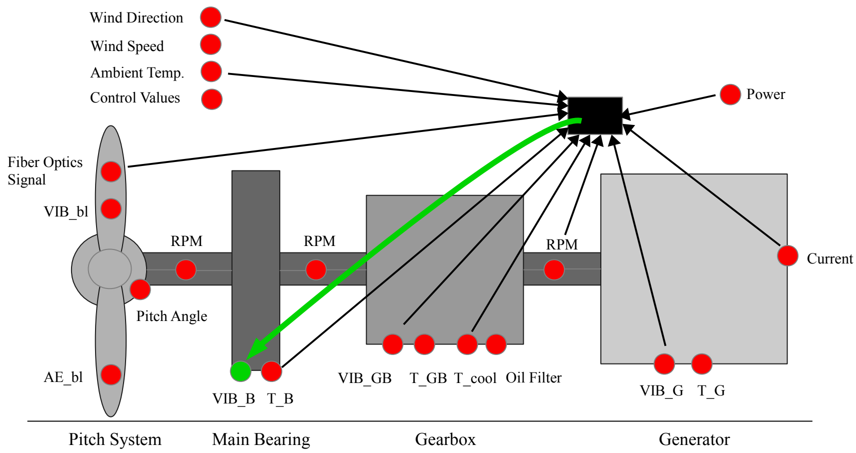


Fig. 1. Exemplary scheme for modelling the main bearing vibration (VIB_B, green dot) with the conversion function (black box). The input signals for the conversion function are depicted with red dots. T: temperature, VIB: vibration, RPM: rotational speed, AE: acoustic emission, bl: blade, B: main bearing, GB: gearbox, G: generator.

reasonable to further normalise the temperatures, speeds or or oil particle counts, calculating amplitudes in characteristic frequency bands for accelerations or use amplitudes or energy measures for acoustic emission.

In the second step, conversion functions are trained with measured data. The conversion function f needs to be capable of modelling non-linear relationships. A suitable regression technique has to be selected according to the properties of the signal as techniques might perform differently for predicting different kinds of signals. The selected training window must be representative for the behaviour of the turbine and sufficiently long.

In the final step, the conversion functions are evaluated in terms of the prediction accuracy. If all possible inputs are used then the prediction performance can be assessed to determine whether a signal is independent. If it is representable by a function of the other signals, the contribution of the individual inputs has to be evaluated with a sensitivity study to identify hidden relationships.

3. Case study

A brief case study with only SCADA data is conducted to test the proposed procedure. Although, the intention is to combine SCADA and CMS signals for a full analysis, this application to real data demonstrates the methodology. In the absence of a full set of signals, six SCADA sensor signals from an offshore wind turbine provided as 10-minute average are available as listed in Table 1. In this simple case, pre-processing of signals as the step 1 of the procedure can be reduced to a validity check as feature extraction is not required.

Table 1. SCADA signals used in the analysis provided as 10 minute average.

Variable	Unit
Rotor speed (low speed shaft)	rpm
Pitch angle	deg
Yaw angle	deg
Tower-top acceleration in x-direction (fore-aft)	m/s^2
Tower-top acceleration in y-direction (side-side)	m/s^2
Active power	MW

In order to find the most suitable tool for building the conversion functions in step 2, four different regression techniques are tested on two of the signals: (i) Random Forests (RF), (ii) Generalized Linear Model (GLM), (iii) Gradient Boost Machine (GBM) and (iv) Artificial Neural Networks (ANNs).

GLMs are flexible generalised linear regression models, formulated by Nelder & Wedderburn [16]. These types of model allow the error to be a distribution other than the normal distribution. A canonical link function for a gaussian distribution is used. Maximum likelihood estimation is used to fit the model to the data.

GBM is a regression and classification technique that originated from the idea of gradient boosting in order to create an optimisation algorithm on a cost function in Breiman [17]. The regression gradient boosting algorithm was developed by Friedman et al.[18]. GBM builds a stage-wise prediction model that consists of a combination of weak prediction models, such as e.g. decision trees.

RF is a machine learning algorithm, which is also frequently used for classification and regression analysis. The main idea was firstly developed in 1995 by Ho et al. [19] and later Breiman et al. [20] introduced the actual algorithm known as random forests. The latter uses a combination of tree predictors and an out-of-bag error as estimate for the generalisation error. The predictor variable importance is then obtained using permutation.

ANNs are a widely used tool for learning non-linear relationships inspired by the human brain. In a network of layers of nodes, each node's output is fed to all nodes in the next layer. In this study, an architecture of feed-forward network with one input, one hidden and one output layer with 20 neurons in the hidden layer is trained by Levenberg-Marquardt backpropagation [21]. Each neuron in the hidden layer produces an output with a hyperbolic tangent sigmoid transfer function. ANNs have been applied successfully to wind turbine condition monitoring, e.g. [12].

The performance is evaluated for different training times. For this purpose, three datasets, as explained in Table 2, corresponding to cases *(A)*, *(B)* and *(C)* are used. Each set is divided into a training and a testing subset.

Table 2. Definition of training and testing subsets for increasing length of the dataset.

	Case *(A)*	**Case *(B)***	**Case *(C)***
Training	48 days	108 days	156 days
Testing	16 days	36 days	52 days
Total	64 days	144 days	208 days

In order to evaluate the performance of the conversion functions in step 3, the performance metrics root mean square error (RMSE), mean absolute error (MAE) and the coefficient of determination R^2 are used, defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \qquad MAE = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(\hat{y}_i - y_i)^2}, \qquad R^2 = 1 - \frac{\sigma(\hat{y} - y)^2}{\sigma(y)^2}; \tag{2}$$

where $\hat{y}$ is the predicted variable, y its true value and σ denotes the standard deviation.

The sensitivity of the conversion function accuracy to the different inputs is analysed by training and testing of conversion functions for all possible combinations of inputs (31 combinations for 5 possible predictors for each target) based on case *(C)*.

4. Case study results

In this section the results are presented for the comparison of modelling techniques and the input sensitivity study.

Comparison of modelling techniques

Table 3 and Table 4 summarise the blind testing results for the predictions of the active power and the tower acceleration in x-direction using the four different techniques and three different datasets. The values for *MAE* and *RMSE* are all normalised to the rated power of the turbine or the maximum tower acceleration in x-direction, respectively. It can be seen, that the GLM did not perform well in predicting the active power and thus it will not be considered further. RF, GBM and ANNs performed well and showed similar results. An extended training time seemed to influence the performance of RF and GBM positively, resulting in lower values for MAE and RMSE and higher ones for R^2. ANNs showed the best model metrics for case *(B)*. Thus, their performance did not necessarily enhance with higher amounts of input data.

Table 3. Results for the MAE, RMSE and R^2 with different training input sizes for predicting the active power (MAE and RMSE normalised to rated power).

Technique	**Case *(A)***			**Case *(B)***			**Case *(C)***		
	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
GLM	0.139	0.170	0.781	0.160	0.187	0.768	0.167	0.194	0.759
RF	0.021	0.035	0.990	0.017	0.032	0.993	0.016	0.031	0.994
GBM	0.017	0.030	0.993	0.015	0.031	0.994	0.014	0.028	0.995
ANNs	0.021	0.035	0.991	0.019	0.030	0.994	0.020	0.034	0.993

Table 4. Results for the MAE, RMSE and R^2 with different training input sizes for predicting the Tower acceleration in the x-direction (MAE and RMSE normalised to maximum value).

Technique	**Case *(A)***			**Case *(B)***			**Case *(C)***		
	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
GLM	0.194	0.230	0.301	0.210	0.251	0.245	0.207	0.247	0.273
RF	0.103	0.142	0.740	0.091	0.130	0.809	0.091	0.127	0.811
GBM	0.084	0.132	0.790	0.070	0.115	0.851	0.073	0.115	0.850
ANNs	0.050	0.094	0.884	0.039	0.075	0.933	0.054	0.093	0.899

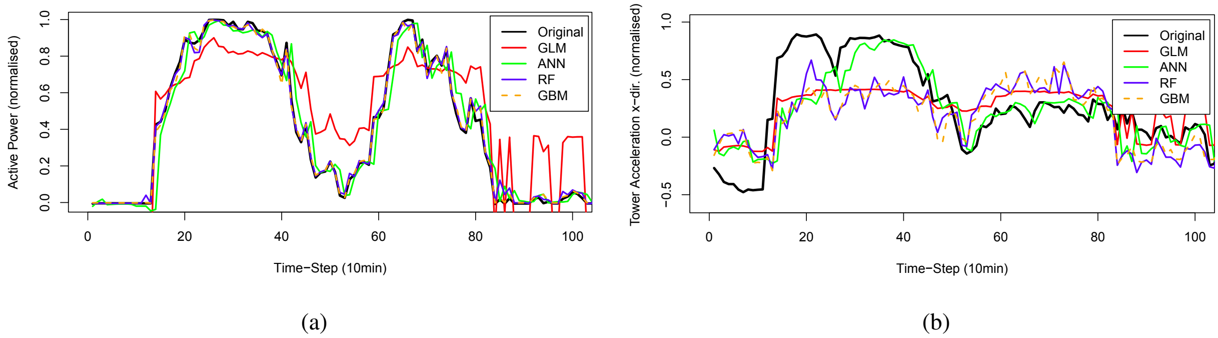
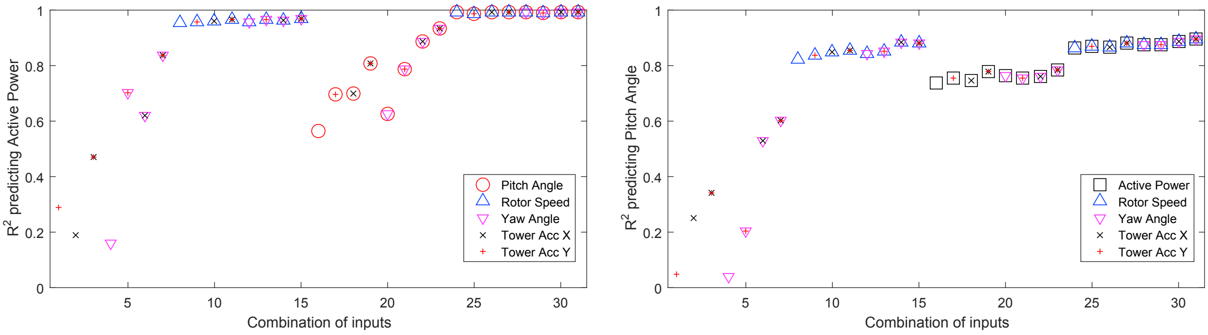


Fig. 2. Original and predicted *(a)* power production and *(b)* tower acceleration in the x-direction for 100 time-steps based on case *(C)*.

As shown in Table 4, for the blind testing of the tower acceleration prediction the ANNs clearly showed the best results for all three cases. Regarding the training time, GBM and ANNs showed both a slightly better performance for input case *(B)*. Figure 2a shows the original and predicted active power using the four different techniques, normalised to the turbine's rated capacity. Figure 2b shows the original and predicted tower acceleration in the x-direction normalised to the maximum value of the original data set. Due to limited space only a selection could be displayed, thus, both figures show the first 100 time-steps of the blind testing based on case *(C)*. The better prediction of the active power compared to the acceleration is due to the strong correlation of the active power to all variables employed in the analysis. For the tower acceleration this is not the case, as discussed in the subsequent sensitivity analysis.

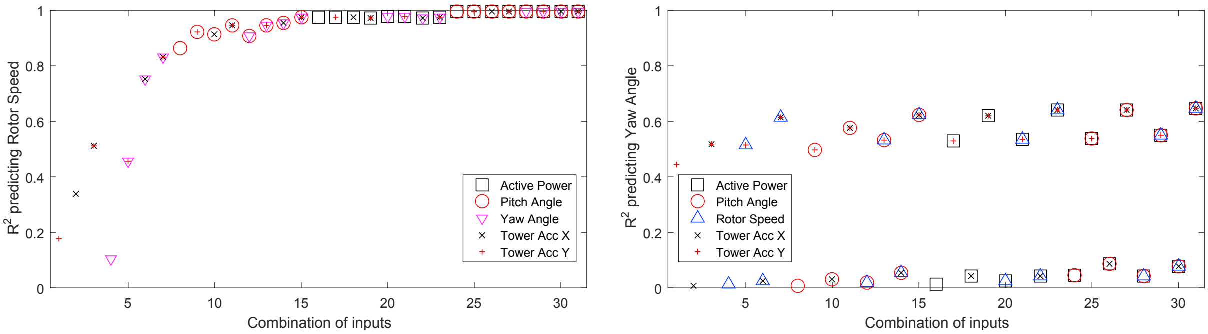
Input sensitivity study

As ANNs performed well for both the prediction of the active power and the tower accelerations, the sensitivity study on the variable importance is carried out with ANNs. Figures 3 - 5 show the results of the input sensitivity study. As there are five possible inputs for each predicted signal, 31 combinations are given on the abscissa. The performance of each input combination is given as the testing R^2 on the ordinate based on case *(C)*. The composition of each combination is displayed by overlaying different markers for the inputs. It can be seen that rotor speed and pitch angle are the most important signals for predicting the active power. If these two are combined, adding



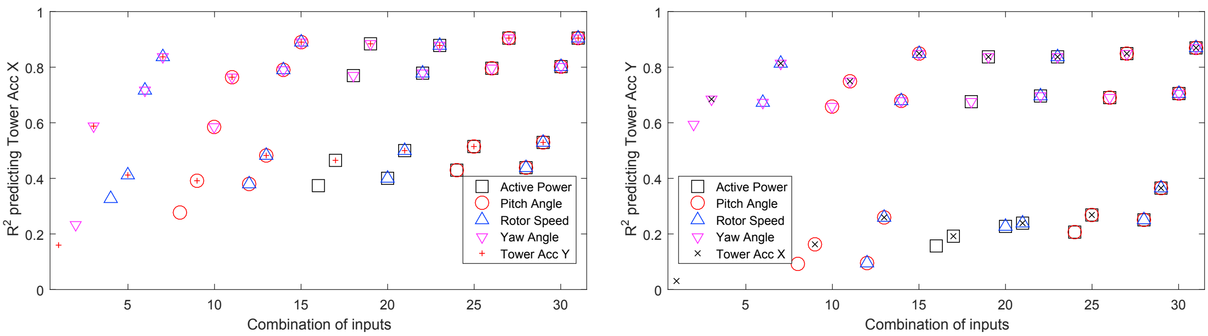
(a) (b)

Fig. 3. Modelling accuracy for all possible input combinations if predicting *(a)* active power and *(b)* pitch angle.



(a) (b)

Fig. 4. Modelling accuracy for all possible input combinations if predicting *(a)* rotor speed and *(b)* yaw angle.



(a) (b)

Fig. 5. Modelling accuracy for all possible input combinations if predicting *(a)* tower-x acceleration and *(b)* tower-y acceleration.

the remaining signals does not contribute to a significantly better accuracy. Similarly, only active power and rotor speed are necessary to predict the pitch angle. The active power is also the best single input predictor for the rotor speed. Surprisingly, predicting the yaw angle is possible if the tower acceleration in the y-direction is included in the inputs. This might reveal the presence of site-specific effects of misalignments causing this relationship. The tower acceleration in the x-direction is best modelled in all cases by a single input if again active power is chosen. However, adding yaw angle as an input is a clear benefit in all cases. The strong relationship of tower acceleration in the y-direction and yaw angle can also be seen if the tower y acceleration is predicted.

Figure 6 visualises the identified relationships of the SCADA signals. An arrow from one node to another stands for a prediction of the target node signal using the start node signal. The thickness of the connections between the

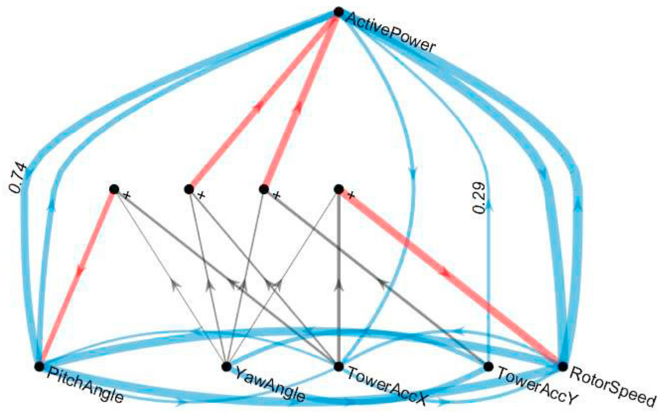


Fig. 6. Diagram of the relationship between investigated SCADA signals in terms of correlation measure R^2. Blue arrows depict single-input predictions (with $R^2 > 0.25$ for clarity), grey arrows contribute to a combination of two inputs in a node marked with + and red arrows combined predictions significantly better than individual modelling.

nodes depicts the accuracy of prediction in terms of R^2. The strong relationship of active power, pitch angle and rotor speed is easily identifiable. Synergistic effects are shown by adding accuracies of predictions with two combined inputs, which are significantly better than predicting with the individual inputs. The strongest synergistic effects are seen in combining yaw angle with the tower accelerations.

5. Conclusions and outlook

In this work a novel approach for wind turbine condition monitoring was presented focusing on synergistic effects between measured signals. The concept is based on predicting certain sensor signals using the information of other sensors at different locations in the turbine. Conversion functions are employed in order to make these predictions. In this work four machine learning algorithms: generalised linear models (GLM), gradient boosting machine (GBM), random forests (RF) and artificial neural networks (ANN) are tested for the conversion functions. Their performance was evaluated in a case study using 10 minute average SCADA data from an offshore wind farm. Three different sizes of the training dataset were used in order to analyse the effects of different training times on the quality of the outcome.

The results show that for predicting each of the parameters the different algorithms performed differently. GBM, RF and ANNs showed very good results for both of the presented predictions. Nonetheless, ANNs showed slightly better results, especially for predicting the tower acceleration, and were used to carry out a sensitivity study demonstrating the variable importance of the predictors and the predicted parameters. The sensitivity study suggests how to interpret the synergistic effects of combined measurements to predict a specific response.

The presented approach has shown to work well and will be extended in future studies. As the case study only contained a limited sample data set, the machine learning algorithms will have to be tested on different datasets and a sensitivity study has to be carried out for the contained parameters. These can contain for instance CMS data in combination with SCADA data. The challenge here will be to understand how signals can be analysed at different sampling frequencies, time windows and continuity in time. Also, the optimal training time should be investigated, as it has been stated that especially ANNs are very sensitive to the latter and do not necessarily show better results with more input data. Furthermore, extending the conversion functions so that measurements from one wind turbine are used to model the status of other turbines in the wind farm, could reduce monitoring costs even further.

In conclusion, further development of this novel approach may result in several benefits for wind farm O&M practice. These range from possible economic benefits by omitting costly sensors in condition monitoring systems, as their information could be predicted using other measurements, to controlling the accuracy of certain sensor outputs by modelling their supposed thresholds. Nevertheless, additional research is also needed to quantify the real economic benefit of O&M by the method proposed.

Acknowledgements

The authors thank the participants of the 1^{st} *Joint Industrial Workshop* (JIW) within the AWESOME project, in particular Estefania Artigao, Ravi Pandit, Lisa Ziegler, Michael Muskulus and Ursula Smolka, who contributed to the development of the idea.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 642108.

References

[1] Lu, B., Li, Y., Wu, X., Yang, Z.. A review of recent advances in wind turbine condition monitoring and fault diagnosis. 2009. In: *2009 IEEE Power Electronics and Machines in Wind Applications*;(1):1-7.

[2] Rademakers, L., Braam, H., Obdam, T., Pieterman, R.. Operation and maintenance cost estimator (OMCE) to estimate the future O&M costs of offshore wind farms. *European Offshore Wind 2009 Conference* 2009;(1):14–16.

[3] Tchakoua, P., Wamkeue, R., Ouhrouche, M., Slaoui-Hasnaoui, F., Tameghe, T., Ekemb, G.. Wind Turbine Condition Monitoring: State-of-the-Art Review, New Trends, and Future Challenges. *Energies* 2014;7(4):2595–2630.

[4] Carroll, J., McDonald, A., McMillan, D.. Failure rate, repair time and unscheduled O&M cost analysis of offshore wind turbines. *Wind Energy* 2016;19(6):1107–1119.

[5] ReliaWind, . Project final report - RELIAWIND - Reliability - Focused Research on Optimizing Wind Energy System Design, Operation and Maintenance: Tools, Proof of Concepts, Guidelines & Methodologies for a new Generation Funding. Tech. Rep.; 2011. URL: *http://cordis.europa.eu/result/rcn/55560_es.html*. Last Accessed: 12/01/2017.

[6] Reder, M.D., Gonzalez, E., Melero, J.J.. Wind Turbine Failures - Tackling current Problems in Failure Data Analysis. *Journal of Physics: Conference Series* 2016;753(072027):1–11.

[7] Randall, R.B.. *Vibration-based Condition Monitoring: Industrial, Aerospace and Automotive Applications*. Chichester, UK: John Wiley & Sons, Ltd; 2011.

[8] Davies, A., editor. *Handbook of Condition Monitoring*. Dordrecht: Springer Netherlands; 1998.

[9] Nilsson, J., Bertling, L.. Maintenance Management of Wind Power Systems Using Condition Monitoring Systems& Life Cycle Cost Analysis for Two Case Studies. *IEEE Transactions on Energy Conversion* 2007;22(1):223–229.

[10] Besnard, F., Nilsson, J., Bertling, L.. On the economic benefits of using condition monitoring systems for maintenance management of wind power systems. 2010. In: *IEEE 11th International Conference on Probabilistic Methods Applied to Power Systems*;(1):160–165.

[11] Kusiak, A., Verma, A.. A Data-Driven Approach for Monitoring Blade Pitch Faults in Wind Turbines. *IEEE Transactions on Sustainable Energy* 2010;2(1):87–96.

[12] Bangalore, P., Tjernberg, L.B.. An Artificial Neural Network Approach for Early Fault Detection of Gearbox Bearings. *IEEE Transactions on Smart Grid* 2015;6(2):980–987.

[13] Tautz-Weinert, J., Watson, S.J.. Using SCADA data for wind turbine condition monitoring - a review. *IET Renewable Power Generation* 2016;In Press. URL: *http://dx.doi.org/10.1049/iet-rpg.2016.0248*. Last Accessed: 04/04/2017.

[14] European Technology and Innovation Platform on Wind Energy, . Strategic research and innovation agenda 2016. Tech. Rep. September; 2016. URL: *https://etipwind.eu/*. Last Accessed: 12/01/2017.

[15] Artigao, E., Colone, L., Pandit, R., Reder, M., Weinert, J., Ziegler, L.. Optimisation of data acquisition in wind turbines with data-driven conversion functions for measurements. Tech. Rep.; 2016. Melero, J.J., Muskulus, M., Smolka, U., editors. In: *1 st Joint Industry Workshop Scientific report*; Zaragoza, Spain; URL: *http://awesome-h2020.eu/1st-joint-industry-workshop-scientific-report/*. Last Accessed: 12/01/2017.

[16] Nelder, J., Wedderbu, R.. Generalized Linear Models. *Journal of the Royal Statistical Society, Series A-General* 1972;135(3):370–384.

[17] Breiman, L.. Arcing the edge. Tech. Rep.; Statistics Department - University of California; Berkeley, CA; 1997. URL: *https://www.stat.berkeley.edu/ breiman/arcing-the-edge.pdf*. Last Accessed: 12/01/2017.

[18] Friedman, J.H... Greedy Function Approximation : A Gradient Boosting Machine. *The Annals of Statistics* 2001;29(5):1189–1232.

[19] Ho, T.K.. Random Decision Forests. 1995. In: *Proceedings of the 3rd International Conference on Document Analysis and Recognition*;1:278–282.

[20] Breiman, L.. Random Forests. Tech. Rep.; Statistics Department - University of California; Berkeley, CA; 2001. URL: *http://link.springer.com/10.1023/A:1010933404324*. Last Accessed: 12/01/2017.

[21] Marquardt, D.W.. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics* 1963;11(2):431–441.