Anomaly Detection Indicators of a Wind Turbine Gearbox Based on Feature Extraction from its Vibration Performance

Miguel Martínez Montaña[1], Miguel A. Sanz-Bobi[2]

[1,2]*Comillas Pontifical University, Engineering School Institute for Research in Technology, Madrid, 28015, Spain*
masanz@comillas.edu

ABSTRACT

This paper proposes a method for obtaining several health condition indicators for wind turbines based on vibration data driven from two similar experimental turbines (damaged and healthy). These indicators are able to capture the bearing and gear condition of the gearbox in the wind turbines. Signal processing and feature extraction were carried out –on both the time and frequency domains– from raw data in order to generate datasets for each shaft of power of the wind turbines. Based on good health condition data, a data mining approach was used to build two reference models for the indicators, one using Self-Organizing Maps (SOM) and another one using Gaussian Mixture Models (GMM). These reference patterns for the indicators were tested with a dataset coming from a damaged wind turbine and the results obtained confirmed the adequacy of these indicators to detect anomalies in the health condition of a wind turbine.

1. INTRODUCTION

Maintenance costs are an important point to take into account in renewable energies in general, and in wind electrical generation in particular (IRENA, 2012). Current wind turbines are designed to work around 120 000 hours a year, over a 20-year lifetime. On average, associated costs with maintenance imply a proportion of the total investment of approximately 3%.

Although both failure type and nature range widely, mechanical wind turbine damage has greater incidence during downtime periods than any other fault type (Spinato, 2009). The basic requirements for wind turbine supervision have been usually formulated (DNVGL, 2016) based on their rolling elements condition (drivetrain, gears and bearings). In fact, these are considered to be representative enough of the wind turbine health condition. In the area of vibration monitoring, the present wind industry is focused on predictive vibration-based maintenance, whose usefulness has been experimentally proven and has turned into standard turbine equipment.

Currently, the state-of-art of vibration signal characterization covers a wide number of techniques, on both time domain (statistical parameters, synchronous averaging) and frequency domain (Fourier Transforms, order analysis, envelope analysis, cepstrum, etc), which both turn out to be particularly useful when predicting bearing and gear damage (Kalista 2015, Hussain 2013, Sasmal 2015). Their application to wind turbines is very extended. On the other hand, some advanced techniques related to predictive maintenance, among other, are based on machine learning algorithms, whose application allows for the prediction of potential faults. These can be summarized into two types: supervised (decision trees, K-Nearest Neighbors, Support Vector Machine, etc) and non-supervised (k-means, GMM, SOM, etc), all of which have been broadly used in wind turbine condition monitoring and prognostics (Coronado 2015, Tchakoua 2014). This paper serves as a contribution in this field proposing a methodology for the estimation of two health condition indicators of a wind turbine gearbox based on its vibrational performance.

The paper is organized as follows: Section 2 describes the methodology proposed. Section 3 presents the source of information used in the paper. Section 4 shows the procedure followed to transform the raw data collected in a more convenient data sets for further analysis. Section 5 describes the elaborations of normal behavior models for the case of a healthy gearbox and the two indicators elaborated based on the models. Section 6 tests the indicators with data sets from a damaged wind turbine. Finally, section 7 presents the more relevant conclusions reached.

2. METHODOLOGY PROPOSED

This section describes the method proposed in this paper for monitoring the health condition of the gearbox of a wind turbine based on its vibration measurements at key points of its components where different accelerometers were installed. This objective has been studied by several authors in different papers of scientific literature as for example Nie and Wang (2013) or Xie et al. (2015). However, this paper is an additional contribution in the field considering different perspectives on the use of algorithms from the knowledge discovering area. These algorithms will supply indicators about the health condition of the wind turbine gearbox. Their application will be discussed in both healthy and damaged cases of similar wind turbine gearboxes.

Figure 1 presents a scheme of the method proposed. First, the vibration raw data set from the components of a healthy wind turbine gearbox was processed extracting several features. This was followed by a principal component analysis able to present the most significant features to take into account and the configuration of a training set. Using the training data set, two algorithms were applied to obtain indicators of the gearbox health condition: Self-Organized Maps (SOM) (Kohonen, 2001) and Gaussian Mixture Models (GMM) (Bishop, 2006).

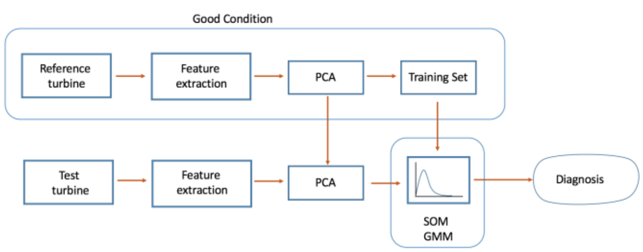

Figure 1. Scheme of the method proposed.

Once the characterization of the normal healthy behavior was done, the same process was followed with data coming from components of a damaged wind turbine gearbox and a comparison of results was carried out in order to demonstrate the ability of the indicators to monitor the health condition of the gearbox. Monitoring these indicators is useful to reduce the time of fault discovering, re-schedule the planned maintenance planning or to suggest a deeper investigation with advanced diagnosis tools.

3. CHARACTERISTICS OF THE CASE STUDIED

The method proposed in the paper was developed using as an application a data collection provided by the National Renewable Energy Laboratory (NREL) (Sheng, 2013). This data set includes vibrations coming from two 750 kW three-bladed identical wind turbines composed of a main low speed shaft, gearbox, high speed shaft and the electrical generator. The data set is coming from different accelerometers located at several points of the gearboxes, whose internal configuration is divided into four shafts: Planet Carrier (PLC), Low-Speed Shaft (LS), Intermediate-Speed Shaft (IMS) and High-Speed Shaft (HS) and three stages: low, intermediate and high speed stage. Figure 2 shows the four shafts in horizontal alignment and the three stages in vertical groups of boxes that correspond to the same stage.

Also Figure 2 shows the location of 8 accelerometers named ANX (X: 1 to 8) where vibration data are collected in both wind turbines. There are two similar 10 minute tests, one coming from the healthy wind turbine and the other from the damaged one. A detailed description of all the gearbox components (gears, bearings), accelerometers, test conditions and turbine faults can be found in Sheng (2013). According to the information about the physical characteristics of the gearbox, the expected characteristic frequencies were estimated obtaining:

- 4 values corresponding to the frequencies of each of the four shafts (SF, Shaft Frequency))
- 3 values corresponding to the frequencies of each stage in the gearbox (GMF, Gear Mesh Frequency)
- 13x4 values corresponding to 13 bearings of the gearbox and their following 4 characteristic frequencies: Outer Race Frequency (BPO), Inner Race Frequency (BPI), Ball Spin Frequency (BS) and Fundamental Train Frequency (FT). The 13 bearings are represented in Figure 2 by triangles.

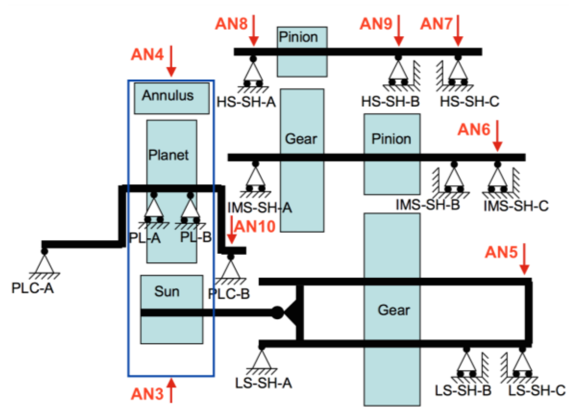

Figure 2. Scheme of the gearbox studied and location of the vibration sensors.

4. PRE-PROCESSING AND FEATURE EXTRACTION

The goal of the first step of the methodology proposed is to prepare the data collected from different accelerometers located in key points of the wind turbine gearbox in order to be useful as inputs for feeding algorithms in charge of estimating the health condition of the gearbox. This first step of the methodology includes two sequential tasks: pre-processing of the raw data and extraction of relevant features for the diagnosis of the gearbox components. Both will be described in the next sub-sections.

4.1. Pre-processing of the raw vibration data

The raw data collected by the 8 accelerometers during the test period was filtered and pre-processed in order to reduce the background noise of the data collected and also to obtain a set of relevant characteristics to be thoroughly analyzed in a

later step. The first task was the elaboration of signal periods of equal lengths on each one of the available signals. After that the use of *Time Synchronous Averaging* (TSA) was initially studied as a first option for data analysis. However, several issues were encountered such as the short time length of the test, odd gear ratios and a lack of tachometer from some tests. Thus, according to MacFadden (2000) the TSA is equivalent to applying a *Comb type* filter in the frequency domain and for this reason *Comb Filters* were finally used to filter out the SF and GMF (Gear Mesh Frequency) harmonics from the data set. Additionally, an envelope analysis was carried out oriented to the bearing diagnosis by means of a spectral kurtosis filtering. A kurtogram implementation (Antoni, 2007 and Wang, 2013) was useful to determine those frequency bands which were most likely to have BPI, BPO and BS impulses coming from the bearings.
Figure 3 presents the scheme followed for processing each signal coming from the accelerometers. As a result of the pre-processing procedure, three sub-signals were obtained x_{GMF}, d_{GMF} (GMF filtered and residual) and its harmonics, and e_{BF} which is the filtered envelope at an optimum band given by a kurtogram. Figure 3 shows the way to obtain all of these sub-signals.

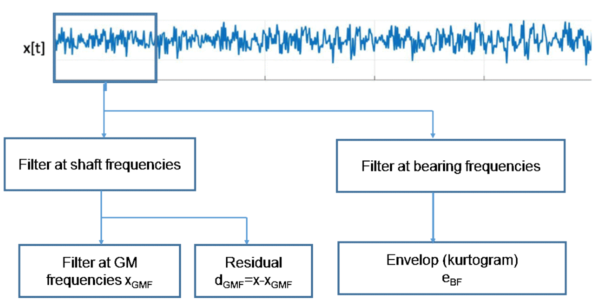

Figure 3. Raw signal processing.

4.2. Feature extraction

After pre-processing the raw data coming from the 8 accelerometers mentioned in section 3, several features were extracted in order to characterize the information collected and already filtered relative to the behavior of the gearbox components. Scientific literature suggests a wide variety of possible features to be extracted (Qiao, 2015, Xie, 2015 and Wang, 2017) from vibrational data of a wind turbine gearbox. In this analysis thirteen features were estimated and tested in total, seven from the sub-signal previously obtained x_{GMF} (combined with d_{GMF}) and six from the e_{BF} respectively. The list of these features is presented in Table 2.

Features #1, 2, 3 and #8, 9, 10 are time domain statistical parameters, while features #5, 6, 7 are condition indicators widely used to detect gear mesh defects.

Features #4,11,12,13 were based on the average FFT, windowed by a Hamming function with a 50% overlap. Equation 1 describes the procedure followed for the estimation of PR for the signal X. It is able to measure the energy spectrum frequency peaks.

$$PR_X[f] = \frac{\sum_{\Delta f_1} X[f]^2}{\sum_{\Delta f_2} R[f]^2},\ \Delta f_1 = 0.03f,\ \Delta f_2 = \alpha \Delta f_1 \tag{1}$$

where $X[f]$ and $R[f]$ represent the original signal and the residual signal respectively at frequency f, after removing all frequency magnitudes contained in Δf_1. After several trials the parameter α was set at 5 and the spectral resolution was adjusted as best possible. Finally $\pm$4Hz was selected with 600 signal portions.

Table 2. List of extracted features.

	#	Name	Equation
Gear diagnosis	1	Root mean square (rms)	$\sqrt{\sum_{t=1}^{N} x_{GMF}[t]^2}$
	2	Kurtosis ($kurt$)	$\frac{m_4(x_{GMF})}{m_2(x_{GMF})^2}$
	3	Crest Factor (CF)	$\frac{\max(x_{GMF}) - \min(x_{GMF})}{rms(x_{GMF})}$
	4	Spectral Power at GMF (PS_{GMF})	$10\log\left(\sum_{k=1}^{K} PR_{GMF}[k]\right)\ \forall k\ GMF\ multiple$
	5	$FM0$	$\frac{\max(x_{GMF}) - \min(x_{GMF})}{\sum_1^K X_{GMF}[k]}\ \forall k\ GMF\ multiple$
	6	$FM4$	$\frac{m_4(d_{GMF})}{m_2(d_{GMF})^2}$
	7	Energy Ratio (ER)	$\frac{rms(d_{GMF})}{rms(x_{GMF})}$
Bearing diagnosis	8	Fisher Skewedness (sk)	$\frac{m_3(e_{BF})}{\sigma(e_{BF})^3}$
	9	Kurtosis ($kurt$)	$\frac{m_4(e_{BF})}{m_2(e_{BF})^2}$
	10	Crest Factor (CF)	$\frac{\max(e_{BF}) - \min(e_{BF})}{rms(e_{BF})}$
	11	Spectral Power at BPI (PS_{BPI})	$10\log\left(\sum_{k=1}^{K} PR_{BPI}[k]\right)\ \forall k\ BPI\ multiple$
	12	Spectral Power at BPO (PS_{BPO})	$10\log\left(\sum_{k=1}^{K} PR_{BPO}[k]\right)\ \forall k\ BPO\ multiple$
	13	Spectral Power at BS (PS_{BS})	$10\log\left(\sum_{k=1}^{K} PR_{BS}[k]\right)\ \forall k\ BS\ multiple$

$m_i(x) = \frac{1}{N}\sum_{t=1}^{N}(x[t]-\bar{x})^i$ is i-th central moment, $\sigma(x) = \sqrt{\frac{N}{N-1}m_2(x)}$ the standard deviation; SF, GMF, BPI, BPO, BS represent shaft, gear mesh, inner defect, outer defect and rolling element frequency respectively; $X[k]$ is the FFT amplitude of $x[t]$ at frequency k. PR is described below.

5. ESTIMATION OF INDICATORS FOR THE WIND TURBINE GEARBOX CONDITION

This section describes two indicators useful to know the health condition of a wind turbine gearbox. They are based on a similar approach, but using different algorithms. The idea behind the indicators is to present if the current vibrational behavior observed corresponds or not to a normal condition. In order to do this the indicators are modelled previously with data sets where no anomalies are present. After that they can be used for observing new information previously not seen. In the next sub-sections the process to select the more relevant features is described in order to configure the training sets and after that two different algorithms are applied in order to model the indicators. The details of this process follow.

5.1. Feature selection

After the pre-processing and feature extraction described in section 4, the information available consists of eight feature datasets, one per sensor, each one composed of 600 observations and 13 variables (features extracted), for each case (healthy and damaged wind turbine gearbox). All data were normalized by means of z-scores, since they had different scales.

The relevance of the 13 features extracted could be different and even null in some cases, and for this reason, a selection of the best features explaining the dynamics of the data was carried out based on a Principal Component Analysis (PCA) (Bishop 2006). The eight data sets were grouped according to the four shafts in the gearbox named PLC, LS, IMS and HS as indicated in Figure 2. Table 3 shows in the column named *sensors* the data sets collected from the accelerometers in Figure 2 that are considered to monitor each shaft. Also each data set was divided in two, one containing the variables or features 1 to 7 that correspond to the gears monitoring and 8 to 13 that correspond to the bearings monitoring. The labels of these two sets are identified in the column *label* in Table 3.

Each dataset was filtered in order to eliminate highly correlated features and finally a PCA was carried out in order to reduce the number of features that are actually relevant in each data set. The reduction obtained is summarized in Table 3. The column *ID* corresponds to the *Initial Dimension*, the column *FD* corresponds to the *Final Dimension* and the rightmost column includes the variance explained by the features finally selected in each data set. For example the data set D11 had initially 21 variables (7 features x 3 sensors) and after the PCA analysis only 9 were necessary, therefore explaining more than 80 % of the variance.

Table 3. Final training data sets configured.

Shaft	Sensors	Variables per sensor (#)	Label	ID	FD	Explained Variance (%)
PLC	AN3, AN4, AN10	1-7	D11	21	9	80.32
		8-13	D12	18	9	84.18
LS	AN5	1-7	D21	7	3	82.92
		8-13	D22	6	5	92.23
IMS	AN6	1-7	D31	7	3	88.47
		8-13	D32	6	4	87.26
HS	AN7, AN8, AN9	1-7	D41	21	9	83.13
		8-13	D42	18	10	80.65

5.2. SOM-based model

A self-organizing map (SOM) was trained for each dataset of the healthy gearbox (Kohonen 2014). A SOM is a type of artificial neural network trained using unsupervised learning to produce a low-dimensional discretized representation of the input space of the training samples, called a map. The map is represented by neurons which weights are patterns discovered in the learning process that represent certain areas of the input domain. The number of neurons (K) required in the maps was estimated by running a k-means algorithm testing a range from 2 to 100 clusters, obtaining the total sum of squared errors with respect to each centroid (SSE) in each trial. The number of neurons finally selected for each map was based on the elbow rule method. The results are listed in Table 4.

Table 4. SOM dimensions for each data set.

Dataset	D11	D12	D21	D22	D31	D32	D41	D42
K	12	15	25	17	16	15	11	12
SOM Dimensions	4x3	4x4	5x5	4x4	4x4	4x4	4x3	4x3

A health condition indicator was proposed using the BMU (Best Matching Unit) distance computed on each observation, divided by the maximum Euclidean distance among all nodes, as presented in Equation 2.

$$CI_{SOM} = \frac{dist(x, w_{BMU})}{\max_{ij}\{dist(w_i, w_j)\}} \tag{2}$$

In order to detect anomalies based on the indicator estimated in Equation 2 (CI_{SOM}), its distributions were studied in the case of a healthy gearbox for each training data set and a Kolmogorov-Smirnov test was carried out on the distribution of the variable $\log(CI_{SOM})$, checking out a normality tendency with the signification of α=0.05. Finally, a threshold was defined at 95% of fitted distribution, so that each observation outside this limit was considered potentially anomalous or corresponding to a symptom of non-healthy condition.

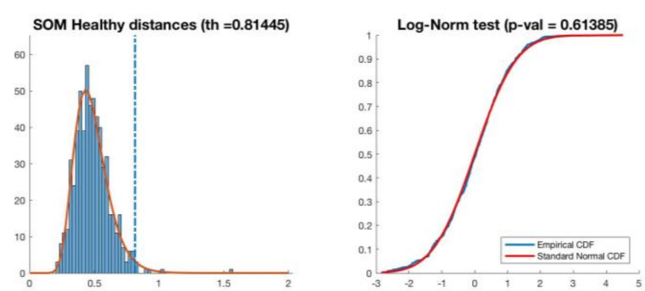

Figure 4. Log Normal test on dataset D41.

Figure 4 presents as an example of the application of this procedure to the data set D41 for the healthy gearbox. In this figure, the distribution of CI_{SOM} values and its fitting by a log-normal function are presented. Also, the values for deciding the threshold and the corresponding p-value are presented. Table 5 summarizes the p-values and thresholds obtained for each training set.

Table 5. Resulting p-values and thresholds from log normal test.

Dataset	D11	D12	D21	D22	D31	D32	D41	D42
p-val	0.124	0.810	0.112	0.809	0.059	0.119	0.614	0.556
p_{95}	1.142	0.326	0.322	0.697	0.137	0.252	0.814	0.467

5.3. GMM-based model

The SOM model obtained previously reaches the goal of a good characterization of the gearbox condition that can be used as a reference for the detection of anomalies in the gearbox components. However in order to robust this result even more, another model was developed using the same data sets but with a different algorithm. In particular, a Gaussian Mixture Model (GMM) was built for each dataset corresponding to healthy condition of the gearbox (Shalev-Shwartz, 2014). A GMM is a probabilistic model that assumes all the samples are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. The goal of these models was to measure how likely new observations belong or not to a previously defined pattern of probability. The GMM models were fitted using an expectation–maximization (EM) algorithm. In order to decide the number of Gaussian distributions to include in each GMM model, three criteria were tested. They were the BIC (Bayesian Information Criterion), the AIC (Akaike Information Criterion) and the K-fold cross validation criterions with a 12.5% data fraction to avoid overfitting. A GMM model was created for each data set testing from 1 to 20 Gaussian components for the three criteria before mentioned.

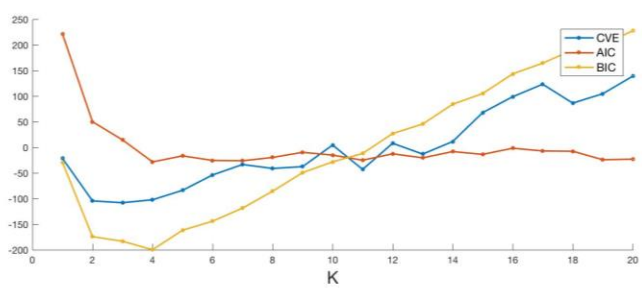

Figure 5. GMM test performance using 1 and 20 Gaussian components using the dataset D21.

Figure 5 shows an example for the case of the data set D21. Finally the BIC criterion was selected because it penalizes more the excess of parameters. Once this selection was done, a final version of the GMM models was obtained using 10% of data fraction. They were validated by computing the ratio of observations included within 95% of the density region ($R_{95\%}$).Table 6 shows the parameters of the GMM models for each data set.

Table 6. GMM parameters for each dataset.

Dataset	D11	D12	D21	D22	D31	D32	D41	D42
K	1	3	3	1	4	2	2	3
$Test\ \log\hat{\mathcal{L}}$	926.081	879.444	328.653	449.869	362.916	352.528	902.851	908.157
$R_{95\%}$	0.927	0.935	0.958	0.950	0.965	0.942	0.945	0.931

According to the models developed a new health indicator CI_{GMM} was estimated based on the density function obtained p(x), according to Equation 3. This indicator takes into account that distant data points from the optimum behavior will have low values of p(x), so CI_{GMM} will increase

$$CI_{GMM} = -\log\left(\frac{p(x)}{max_x\{p(x)\}}\right) \tag{3}$$

6. USE OF THE HEALTH CONDITION INDICATORS FOR DETECTION OF ANOMALIES

The health condition indicators proposed CI_{SOM} and CI_{GMM} were estimated for the datasets of the damaged wind turbine gearbox using the SOM and GMM models obtained from the datasets of the healthy wind turbine gearbox. Thus, for each shaft, the condition of gears and bearings was evaluated throughout the entire 10-minute test.

Figure 6 shows their temporal evolution corresponding to the PLC and HS shafts used as examples. The sub-figures *a* and *c* correspond to the gears and the sub-figures *b* and *d* to the bearings in the shafts PLC and HS respectively. The first result from this figure is that both indicators have similar behaviors and both are able to detect abnormal behavior. Also, it can be observed in Figure 6 that the cases *a*, *c* and *d* present abnormal behavior detected by the indicators in the data sets of the damaged gearbox because they are overpassing the corresponding thresholds of normal behavior expected (discontinue lines). However, in case *b* the indicators are closer to the thresholds of normal behavior and it is more difficult to decide if an anomaly is present. In order to confirm the results shown in Figure 6, a different graphical representation was used facing the indicators corresponding to the gears (horizontal axis) against the indicators corresponding to the bearings (vertical axis) for the four shafts. Only the results for the indicator CI_{GMM} are presented in Figure 7 where in discontinued lines the respective thresholds of normal behavior are indicated, in red dots the data corresponding to the damaged gearbox and in green dots those belonging to the healthy gearbox.

From Figure 7 it is possible to conclude the presence of constant and major faults on gears in the PLC and HS shafts, and also faults in the bearing of the IMS and HS shafts. These conclusions are coherent with the results expected validating the role of the indicators proposed and the methodology exposed in this paper.

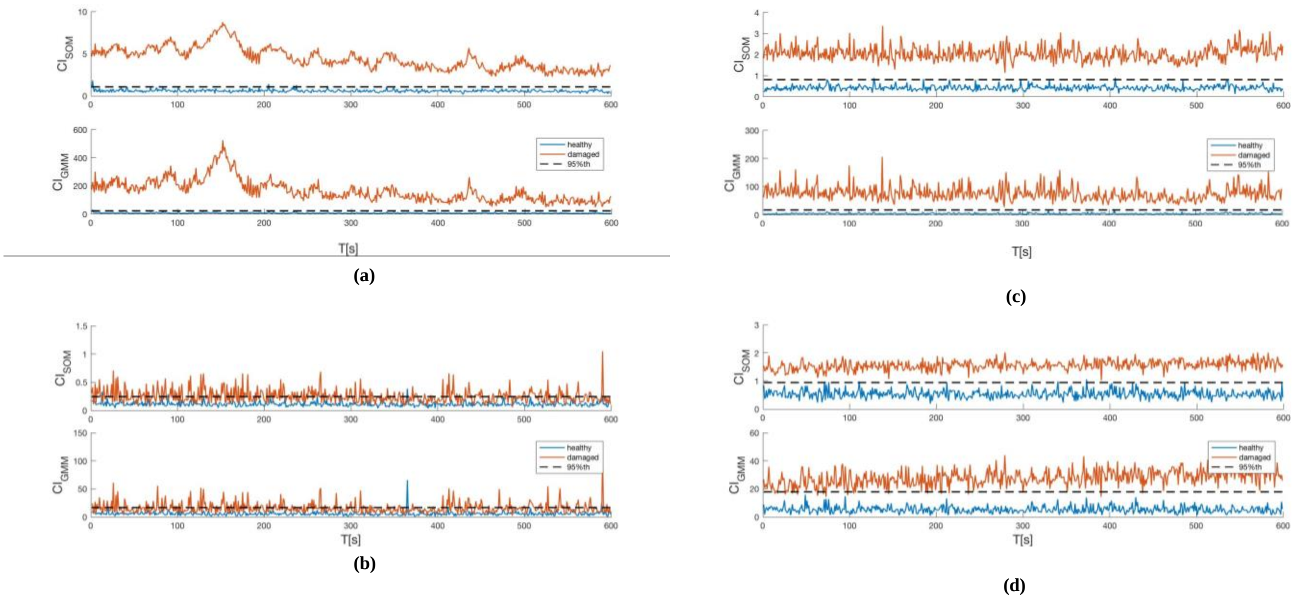

Figure 6. Evolution of the health indicators obtained from the datasets D11(a), D12(b) (PLC shaft), D41(c), D42(d) (HS shaft).

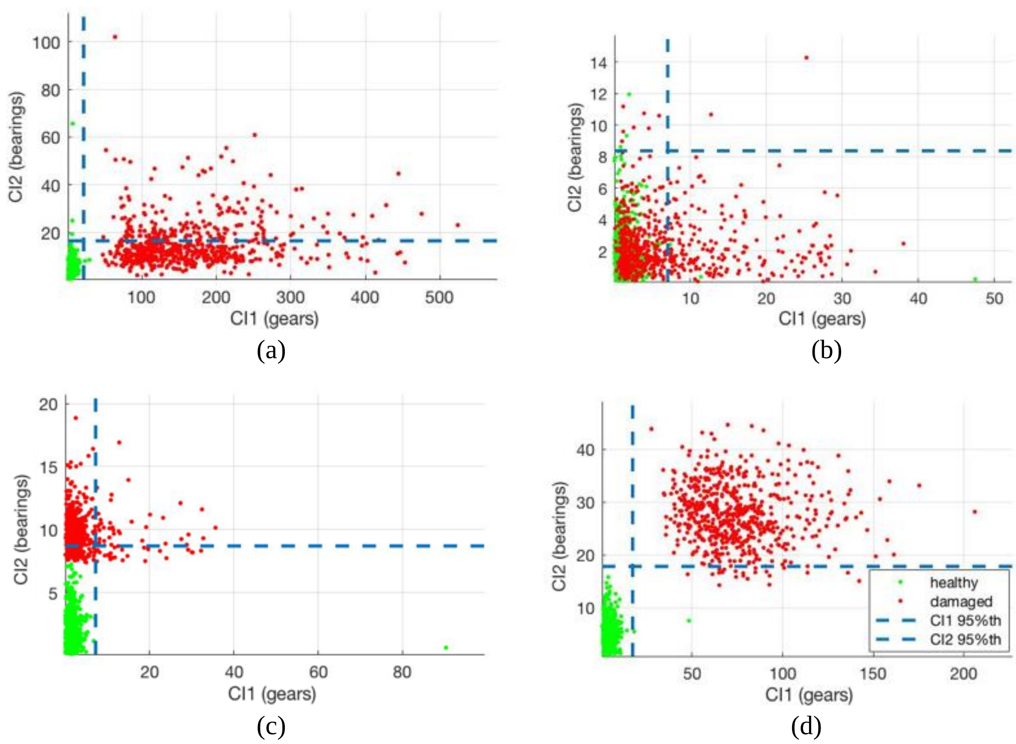

Figure 7. Scatter plots with CI_{GMM} values from each pair of datasets of PLC (a), LS (b), IMS (c) and HS (d).

7. CONCLUSION

This paper has presented a methodology to estimate two indicators able to detect anomalies in a wind turbine gearbox based on vibrations collected from key points of its components. The methodology described covers several steps from the pre-processing of the raw vibrational data until the estimation of normal behavior models and the indicators based on them able to detect anomalies. The two estimated indicators have a similar objective but based on different algorithms, SOM and GMM respectively. The use of them in parallel makes the detection of abnormal behavior more robust.

The models and indicators developed used two different data sets of vibrations collected by accelerometers in two similar gearboxes of a wind turbine, one of then healthy and another one damaged.

The models and indicators obtained demonstrated their ability to detect anomalies in the data sets corresponding to the damaged gearbox suggesting their possible use in the future. However, the proposed method exhibits some limitations such as the best that the available datasets corresponded to a simple point of operation of the gearboxes (1200 rpm in the High Speed shaft) and the collection of data was for a short time period. Also, the combined effect of several failure modes in the damaged data set has to be taken into account. Future studies will try to overpass these limitations. In any case, the usefulness of the proposed method has been proven by different tests as was described in the paper.

ACKNOWLEDGEMENT

The authors are very grateful to the National Renewable Energy Laboratory (NREL), USA for providing the public data used in this paper.

REFERENCES

Antoni, J. (2007) Fast computation of the kurtogram for detection of transient faults, *Mechanical Systems and Signal Processing*, vol 21 no.1, pp. 108-124

Bishop, C. (2006) *Pattern Recognition and Machine Learning*, Springer

Coronado, D., Fischer, K. (2015). Condition Monitoring of Wind Turbines: State of the Art, User Experience and Recommendations. Institute for Wind Energy and Energy System, https://www.vgb.org/vgbmultimedia/383_Final+report-p-9786.pdf (accessed January 2018)

DNV-GL. (2016) Certification of condition monitoring DNVGL-SE-0439. https://rules.dnvgl.com/docs/pdf/DNVGL/SE/2016-06/DNVGL-SE-0439.pdf (accessed January 2018)

Hussain, S. Gabbar, H.A. (2013) Vibration analysis and time series prediction for wind turbine gearbox prognostics, *International Journal of Prognostics and health management*, vol 14(online)

IRENA (2012). Renewable energy technologies: Cost analysis series. Wind power. https://www.irena.org/DocumentDownloads/Publications/RE_Technologies_Cost_Analysis-WIND_POWER.pdf (accessed January 2018)

Kalista, K., Liska, J .(2015). Time-frequency methods for signal analysis in wind turbines. *Journal of Physics: Conference Series 659 012012*

Kohonen,T. (2001) *Self-Organizing Maps*, Springer

Kohonen,T. (2014). *MATLAB Implementations and Applications of the Self-Organizing Map*. Unigrafia Oy, Helsinki, Finland

MacFadden, P.D. Toozhy, M.M. (2000). Application of Time Synchronous Averaging to vibration monitoring of rolling element bearings. *Mechanical Systems and Signal Processing*, vol 14, pp. 891-906

Nie, N., Wang L. (2013) Review of Condition Monitoring and Fault Diagnosis Technologies for Wind Turbine Gearbox, *Procedia CIRP*, vol 11 pp. 287-290

Qiao, W., Lu, D. (2015). A survey on wind turbine condition monitoring and fault diagnosis—Part I: Components and subsystems. *IEEE Transactions on Industrial Electronics*, vol. 62, no. 10, pp. 6536–6545

Santos, P., Villa, L.F., Reñones, A., Bustillo, A., Maudes, J. (2015) An SVM-Based solution for fault detection in wind turbines, *Sensors*, vol 15 pp. 5627-5648. doi: 10.3390/s150305627

Sasmal, J.K. (2015) Condition monitoring of wind turbine gearbox by vibration analysis, *International Journal of Scientific Engineering and Research*, vol 3 no.3 (online)

Shalev-Shwartz, S., Shai Ben-David, S. (2014) *Understanding Machine Learning From Theory to Algorithms*, Cambridge University Press

Sheng, S. (2013). Wind Turbine Gearbox Vibration Condition Monitoring Benchmarking Datasets. National Renewable Energy Laboratory. https://openei.org/datasets/dataset/wind-turbine-gearbox-condition-monitoring-vibration-analysis-benchmarking-datasets

Spinato,F., Tavner, P.J., Van Bussel, G.J.W, Koutoulakos, E. (2009). Reliability of wind turbine subassemblies. *IET Renewable Power Generation*, vol. 3 no. 4, pp. 387 - 401

Tchakoua, P., Wamkeue, R., Ouhrouche, M., Slaoui-Hasnaoui, F., Tameghe, T.A., Ekemb, G. (2014) Wind turbine condition monitoring: state-of-the-art review, new trends, and future challenges, *Energies*, vol 7 pp. 2595-2630. doi: 10.3390/en7042595108-124

Wang, J., Qiao, W., Qu, L. (2017). Wind turbine bearing fault diagnosis based on sparse representation of condition monitoring signals. *Energy Conversion Congress and Exposition (ECCE)*. doi: 10.1109/ECCE.2017.8096654

Wang, D., Tse, P.W., Tsui, K.L. (2013) An enhanced Kurtogram method for fault diagnosis of rolling element bearings, *Mechanical Systems and Signal Processing*, vol 35 no.1-2, pp. 176-199

Xie, Y., Zhifei, G., Ke, T., Mingjie, Z., Yonghai, W. (2015). Wind turbine gearbox condition monitoring system based on vibration signal. *12th IEEE International Conference on Electronic Measurement & Instruments (ICEMI).* doi: 10.1109/ICEMI.2015.7494242

BIOGRAPHIES

Miguel M. Montaña was born in Madrid, in 1993. He received the B.S. degree in Electromechanical Engineering and M.S. in Industrial Engineering both from the Comillas Pontifical University, Madrid (Spain) in 2015 and 2017 respectively. Currently working as logistics engineer, his main interest areas are within statistics, data science and machine learning fields. His Master Thesis research was the origin of this paper.

Miguel A. Sanz-Bobi is Professor at the Telematics and Computer Science Department and also Researcher at the Institute for Research and Technology (IIT) both within the Engineering School of the Comillas Pontifical University, Madrid (Spain). He divides his time between teaching and research in the Artificial Intelligence field applied to diagnosis and maintenance of industrial processes. He has been the main researcher in many industrial projects over the last 25 years related to the diagnosis in real-time of industrial processes, incipient detection of anomalies based on models, knowledge acquisition and representation, data mining, reliability and predictive maintenance. All these projects have been based on a combination of artificial intelligence, new information technologies and data mining techniques. E_mail: masanz@comillas.edu