Renewable Energy 113 (2017) 1490–1502

Contents lists available at ScienceDirect

Renewable Energy

journal homepage: www.elsevier.com/locate/renene

Performance monitoring of a wind turbine using extreme function theory

Evangelos Papatheou [a,*], Nikolaos Dervilis [a], Andrew E. Maguire [b], Carles Campos [b], Ifigeneia Antoniadou [a], Keith Worden [a]

[a] Dynamics Research Group, Department of Mechanical Engineering, University of Sheffield, Mappin Street, Sheffield, S1 3JD, UK
[b] Vattenfall Research & Development, New Renewables, The Tun Building, Holyrood Road, Edinburgh, EH8 8AE, UK

ARTICLE INFO

Article history:
Received 17 November 2016
Received in revised form
1 June 2017
Accepted 3 July 2017
Available online 5 July 2017

Keywords:
Extreme values
Extreme functions
Wind turbines
Power curve monitoring
False-positive rate

ABSTRACT

A power curve relates the power produced by a wind turbine to the wind speed. Usually, such curves are unique to the various types of wind turbines, so that by monitoring the power curves, one may monitor the performance of the turbine itself. Most approaches to monitoring a system or a structure at a basic level, generally aim at differentiating between a normal and an abnormal state. Typically, the normal state is represented by a model, and then abnormal, or extreme data points are identified when they are compared to that model. This comparison is very often done pointwise on scalars in the univariate case, or on vectors, if multivariate features are available. Depending on the actual application, the pointwise approach may be limited, or highly prone to false identifications. This paper presents the use of extreme functions for the performance monitoring of wind turbines. Power curves from an actual wind turbine, are assessed as whole functions, and not individual datapoints, with the help of Gaussian process regression and extreme value distributions, with the ultimate aim of the performance monitoring of the wind turbine at a weekly resolution. The approach is compared to the more conventional pointwise method, and approaches which make use of multivariate features, and is shown to be superior in terms of the number of false identifications, with a significantly lower number of false-positives without sacrificing the sensitivity of the approach.

1. Introduction

Monitoring the health of structures is vital for their safety, as well as beneficial to the reduction of the high costs their maintenance may demand. Wind turbines are becoming increasingly more popular, so that their maintenance is also of high interest. Traditional non-destructive evaluation (NDE) methods [1] are currently the main inspection tool for structures, and they may be an answer to wind turbine monitoring as well. However, although they are highly effective, they have certain disadvantages regarding their use, which include among others, general limitations to accessible areas of the structure, high demand on expertise and also high inspection costs. In addition, NDE methods work in a local vicinity and so usually they necessitate the *a priori* knowledge of the area of interest, and they may disrupt the normal operation of the structure during the inspection process. Therefore, various other monitoring methods have been proposed for wind turbines, from vibration approaches on the blades [2–4] to advanced signal processing in gearboxes [5–7] or bearings [8]. General reviews can be found in Refs. [9,10].

Among the various proposed approaches, which also make use of supervisor control and data acquisition (SCADA) data [11], are the power curves. Wind turbines are designed by manufacturers to have a specific relationship between the power produced and the wind speed, and in general, researchers have exploited the deviation from a reference curve in order to monitor the state of a turbine e.g. Refs. [12,13]. Various ways of modelling a power curve can be found in the literature [14,15]. Other works on using machine learning for modelling power generation can be seen in Refs. [12,16], while in Ref. [17], three operational curves, power, rotor and pitch were used for reference in order to produce control charts [18] for the monitoring of wind turbines. More recently in Ref. [19], wind turbine power curves were modelled with the help

[*] Corresponding author. Now at: University of Exeter, Structures Dynamics and Control Group, College of Engineering Mathematics and Physical Sciences, North Park Road, Exeter, EX4 4QF, UK.
E-mail address: e.papatheou@exeter.ac.uk (E. Papatheou).

http://dx.doi.org/10.1016/j.renene.2017.07.013

of Gaussian processes (GPs) for a full offshore wind farm, and their performance was also monitored with the help of control charts on the residuals of the models.

Most of the previously mentioned methods, including the power curve, can be loosely categorised under the general term of Structural Health Monitoring (SHM), which has emerged as a potential answer to the drawbacks of NDE techniques. Comprehensive reviews on SHM are [20–22]. At the lowest level of SHM, the main objective is simply the detection of the presence of damage. In most cases, a model of normality is built, and data originating from the structure of interest are tested, usually after some processing, in terms of ‘novelty’ (when compared to the normal model). ‘Novel’ data are thus ‘detected’ and can be considered indicative of damage; the approach can thus be termed by some as ‘novelty detection’. Although this process is generally considered less challenging than the full identification of damage i.e. type, location or severity, and it may be under strict laboratory conditions, when it comes to real structures there are still various problems to be addressed.

The majority of novelty detection methods often use some form of statistical test in order to monitor a ‘control’ quantity. Examples may be control charts [18], which are standard applications for online monitoring of such quantities, and they have been used with the power curve method as well [17,19]. In most cases, the identification of the abnormal or extreme values of the ‘control’ quantity is performed pointwise on individual points, and that was also done in Ref. [19]. If multivariate features are used, then usually they are fused in a single scalar quantity, like the Mahalanobis squared-distance [23], which is then tested against a threshold in order to be declared as normal or not. In a probabilistic framework, the tested data are generally assumed as independent and identically distributed (i.i.d.). Depending on the application, this approach may be limited and significantly prone to false identifications (e.g. false alarms), especially for the pointwise comparison, since when it comes to time series, the i.i.d. assumption may not hold. False alarms may be arguably considered a prime reason for industry not adopting SHM/monitoring approaches, as well as can constitute extra economic burden. Conventional efforts to reduce the false alarms often result in significant loss in the sensitivity of the proposed monitoring methodology. Overall, when data from real and complex structures under various loading or environmental conditions are used, then novelty detection may prove challenging.

Recently, a novel approach proposed by Clifton et al. [24], attempts to assess functions instead of single data points with several advantages including improvement in classification results as well as a reduction in false alarms. The functions are represented by time series and can be used to create a model of normality. Subsequently, individual functions can be tested for a single classification decision in terms of their novelty and their extremity - thus the method is named as an Extreme Function Theory (EFT). The purpose of this paper is to apply a modified version of this method, for the first time for SHM, on power curves, constructed from real wind turbine data, in a novelty detection scheme. The power curves are constructed by weekly SCADA data making the approach attractive for industry. The overall approach is compared to a conventional pointwise methodology, as well as approaches which make use of multivariate features.

The layout of the paper is as follows: Section 2 describes the Extreme Function Theory (EFT) starting from an overview, continuing into Gaussian process regression theory - described for the convenience of the reader - before explaining the EFT approach. Section 3 describes the application of the EFT on wind turbine data starting from the description of the data and continuing with a general methodology for the application of EFT on wind turbine data. Section 4 compares the Extreme Function Theory with conventional pointwise approaches, and Section 5 extends the discussion on this comparison. Finally, the paper is rounded off with some conclusions and the overall potential of the approach.

2. Extreme function theory

2.1. Overview

As explained before, the goal of the approach is to classify functions, where the functions are represented by sampled data vectors, and not just the classification of data points. A model of normality is also assumed to be constructed prior to the classification, and the functions have to be tested in terms of their novelty. In a probabilistic approach, one would prefer to assign probabilities to the functions, so a mapping between a test vector and probabilities is required [24]. The Gaussian process (GP) framework provides a convenient probabilistic and non-parametric approach for regression, and is at the heart of the presented methodology [25].

The problem is formulated as follows: given some data sets $\{\mathbf{x_i}, \mathbf{y_i}\}$ where $i = 1 \ldots n$ and n is the number of observations, construct a model of normality M based on the desired training set, then test whether a function consisting of a test set $\{\mathbf{x}^*, \mathbf{y}^*\}$ is extreme compared to the model M. In this approach M will simply be a GP model conditioned on the training set. In the following section the Gaussian process regression will be briefly outlined for the convenience of the reader.

2.2. Gaussian process regression algorithm

Rasmussen and Williams [25] define a Gaussian process (GP) as “a collection of random variables, any finite number of which have a joint Gaussian distribution”. The initial and basic step in order to apply Gaussian process regression is to obtain a mean $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$ as GPs are completely specified by them, $\mathbf{x}$ represents the input vector. So for any real process $f(\mathbf{x})$ one can define,

$$m(\mathbf{x}) = E[f(\mathbf{x})] \tag{1}$$

$$k(\mathbf{x}, \mathbf{x}') = E[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))] \tag{2}$$

where E represents the expectation. Often, for practical reasons because of notation purposes (simplicity) and little knowledge about the data at the initial stage, the prior mean function is set to zero. The Gaussian processes can be defined as,

$$f(\mathbf{x}) \sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{3}$$

If a zero-mean function is assumed, the covariance function can be described as,

$$cov(f(\mathbf{x}_p), f(\mathbf{x}_q)) = k(\mathbf{x}_p, \mathbf{x}_q) = \exp\left(-\frac{1}{2}\left|\mathbf{x}_p - \mathbf{x}_q\right|^2\right) \tag{4}$$

this is the squared-exponential covariance function which is used throughout this paper, for other choices of covariance functions the reader is referred to [25]. Assuming now that one has a set of training outputs $\mathbf{f}$, and a set of test outputs $\mathbf{f}^*$, one has the prior [25],

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}^* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(X,X) & K(X,X^*) \\ K(X^*,X) & K(X^*,X^*) \end{bmatrix}\right) \tag{5}$$

where the capital letters represent matrices. If there are n training points, and n^* testing points, then $K(X,X^*)$ denotes the $n \times n^*$ matrix of the covariances evaluated at all pairs of training and test points, and similarly for the rest of the matrices $K(X,X)$, $K(X^*,X)$ and $K(X^*,X^*)$.

As the prior has been generated by the mean and covariance functions, in order to specify the posterior distribution over the functions, one needs to limit the prior distribution in a such a way that includes only these functions that agree with actual data points. An obvious way to do that is by generating functions from the prior and select only the ones that agree with the actual points. Of course, this is not a realistic way of doing it as it would consume a lot of computational power. In a probabilistic manner this can be done easily via conditioning the joint prior on the observations and this will give (for more details see Refs. [25–27]),

$$\mathbf{f}^* \Big| X^*, X, \mathbf{f} \sim N\left(\begin{array}{l} K(X^*,X)K(X,X)^{-1}\mathbf{f}, \\ K(X^*,X^*) - K(X^*,X)K(X,X)^{-1}K(X,X^*) \end{array}\right) \tag{6}$$

Function values $\mathbf{f}^*$ can be generated by sampling from the joint posterior distribution and at the same time evaluating the mean and covariance matrices from (6).

In practice, it may not be possible to access function values themselves, but rather noisy versions: $\mathbf{y} = f(\mathbf{x}) + \varepsilon$. Assuming i.i.d. Gaussian noise ε with variance σ_n^2, then Equation (6) becomes [25],

$$\mathbf{f}^* | X, \mathbf{y}, X^* \sim N\left(m(\mathbf{f}^*), cov(\mathbf{f}^*)\right), \text{ where} \tag{7}$$

$$m(\mathbf{f}^*) = K(X^*,X)\left[K(X,X) + \sigma_n^2 I\right]^{-1}\mathbf{y} \tag{8}$$

$$cov(\mathbf{f}^*) = K(X^*,X^*) - K(X^*,X)\left[K(X,X) + \sigma_n^2 I\right]^{-1}K(X,X^*) \tag{9}$$

The covariance functions used in this study are usually accompanied by some extra parameters in order to obtain a better control over the types of functions that are considered for the inference. As an example, the squared-exponential covariance function can take the form (1-dimensional),

$$k_y(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{1}{2l^2}(x_p - x_q)^2\right) + \sigma_n^2\,\delta_{pq} \tag{10}$$

where k_y is the covariance for the noisy target set $\{\mathbf{y}\}$. The length-scale l (determines how far one needs to move in input space for the function values to become uncorrelated), the variance σ_f^2 of the signal, and the noisy variance σ_n^2 are free parameters that can be varied. These free parameters are called hyperparameters. The tool that has to be applied for selecting the model for choosing the optimal hyperparameters for GP regression, is the maximum marginal likelihood of the predictions $\mathrm{p}(\mathbf{y}|X,\theta)$ with respect to the hyperparameters θ,

$$\log \mathrm{p}(\mathbf{y}|X,\theta) = -\frac{1}{2}\mathbf{y}^T K_y^{-1}\mathbf{y} - \frac{1}{2}log\left|K_y\right| - \frac{n}{2}\log 2\pi \tag{11}$$

where $K_y = K_f + \sigma_n^2 I$ is the covariance matrix of the noisy test set $\{\mathbf{y}\}$ and K_f is the noise-free covariance matrix. The reader is referred to [25] for the exact solution of the maximisation of the marginal log likelihood through its partial derivatives.

2.3. Extreme function theory with GP regression

Recalling from the previous that the aim of the extreme function theory is to classify a whole test function $\mathbf{f}^*$, with a single decision as extreme or not when compared to a model of normality M (which is based on a Gaussian process), GP regression is exploited [24,25]: the joint distribution over all the test set points $\{\mathbf{x}^*, \mathbf{y}^*\}$ conditioned on the model M is,

$$p(\mathbf{f}^* | \mathbf{x}, \mathbf{f}, \mathbf{x}^*) \sim N(\mu^*, K^*) \tag{12}$$

where $\mathbf{f}$ is the latent function of the GP model M (given by the mean prediction of the GP when provided with $\mathbf{x}$ as an input), and $\mathbf{f}^*$ will be the values of the function to be tested. Based on the general definition of a GP [25], $p(\mathbf{f}^* | \mathbf{x},\ \mathbf{f},\ \mathbf{x}^*)$ should follow a multivariate Gaussian density,

$$p = \frac{1}{\sqrt{(2\pi)^d \left|K^*\right|}} e^{-\frac{1}{2}(\mathbf{f}^* - \mu^*)^T K^{*-1}(\mathbf{f}^* - \mu^*)} \tag{13}$$

where d will be the dimension of the test vector $\mathbf{x}^*$. The mean function μ^* and K^* are given by Equations (8) and (9) once all the hyperparameters are added in (see also [25]),

$$\mu^* = k(\mathbf{x}^*, \mathbf{x})\left[k(\mathbf{x},\mathbf{x}) + \sigma_n^2 I\right]^{-1}\mathbf{y} \tag{14}$$

$$K^* = k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{x})\left[k(\mathbf{x},\mathbf{x}) + \sigma_n^2\right]^{-1}k(\mathbf{x}, \mathbf{x}^*) \tag{15}$$

where $k(\mathbf{x}, \mathbf{x})$ is the chosen covariance function, which in this study will be the squared exponential given by Equation (10).

A probability density z can now be defined as $z = f_n(\mathbf{f}^*)$ given by Equation (13), in order to obtain a single value z expressing the likelihood of a whole test function $\{\mathbf{x}^*, \mathbf{y}^*\}$. The probability density function (pdf) of z will have a corresponding Extreme Value (EV) distribution for low values of z (i.e. in the left tail of the distribution), and it has been shown that it will asymptotically converge to a Weibull distribution function (df), and a closed-form solution for such a df was approximated in Ref. [24].

In this work, in order to calculate appropriate thresholds for z, a slightly different approach to that in Ref. [24] was followed, due to various numerical challenges when the exact method shown in Ref. [24] was attempted. According to Fisher and Tippet [28], when the number of vector samples originating from an arbitrary parent distribution tends to infinity, the induced distribution on the extrema of the samples can only take one of three forms: Gumbel, Weibull, or Frechet. In the case of z, it is expected to converge to a Weibull [24,29] form. In order to estimate the parameters of the Cumulative Distribution Function (CDF), an optimisation algorithm was employed here. The idea is to fit a parametric model to the tails of the parent's distribution, and thus estimate its correct coefficients. After such coefficients are obtained it is trivial to estimate a threshold according to the level of confidence that is desired. The optimisation algorithm employed here was Differential Evolution (DE) [30], and the approach employed was similar to that in Ref. [31]. DE belongs to the family of evolution-based algorithms, where an initial random population of solutions is propagated through a repeated cycle of mutation and crossover operations until an optimal (or near optimal, according to desired criteria) solution is obtained. Inherent in an evolution process is the calculation of a fitness or cost function, which in the particular problem here is the error in a fit of a parametric model to a given cumulative distribution function (CDF). A normalised mean squared error (NMSE) was used for the curve-fit, given by,

$$NMSE(\hat{\mathbf{y}}) = \frac{100}{N\sigma_y^2}\sum_{i=1}^{n}\left(\mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i}\right)^2 \tag{16}$$

where the caret denotes an estimated quantity, y_i is the actual observation, N is the total number of observations and σ_y the standard deviation of y.

3. Application of the extreme function theory on wind turbines: data description and general methodology

3.1. Description of the data and creation of class datasets

The data used in this study are SCADA extracts originating from an actual wind turbine owned by Vattenfall. For confidentiality reasons, no information regarding the actual wind farm, the total number and the type of the wind turbines is disclosed here. The case study presented contains recorded data entirely from one wind turbine in a period of 125 weeks. The mean values of the power produced, and of the measured wind speed, in 10 min intervals, were available.

The main goal is to determine whether a wind turbine is operating in a normal state or not, and in order to do this, the power curve method was used. Examples of a normal (‘good’) and of a ‘bad’ power curve, as provided by Vattenfall, are shown in Fig. 1. The characterisation of the power curves shown in Fig. 1 as ‘good’ or ‘bad’ relies on the expertise of Vattenfall engineers.

In the spirit of using functions to test, and not individual points, the analysis is carried out at a weekly resolution i.e. data from one week correspond to a test function (here, a power curve). It is important to note, that monitoring wind turbines at a weekly resolution has high practical value for a company, and it was also desired by Vattenfall. Moreover, the exercise presented here makes use of wind turbine data (power vs wind speed) without the use of any information on the actual status of the turbines, and this was also desired by Vattenfall (as such information may not be easily available always), and does not aim to assess the sensitivity of the power curve approach to various levels of faults. Since there was no information on actual status, the data were plotted at a weekly resolution and were separated subjectively, but based on informed engineering judgment and using Fig. 1 - which was provided by Vattenfall - as a reference, together with the guidelines as to how one might judge a curve as ‘good’ or ‘bad’. Examples of this class separation can be seen in Fig. 2, where a ‘normal’ week is plotted, and in Fig. 3 where ‘abnormal’ weeks are shown. Figs. 2 and 3 contain weeks which were later identified correctly as ‘normal’ (Fig. 2) and ‘abnormal’ (in Fig. 3) by the proposed EFT approach. Finally, an extra group was also created, one that contained power curves which could not be categorised unambiguously (by visual inspection, employing the criteria provided by Vattenfall experts) either as ‘good’ or as ‘bad’. The latter class could be used as a ‘control’ group, and its correct classification should be crucial for the practical application of monitoring wind turbines at a weekly basis - after all there is a need for a confident and objective decision, rather than just simple visual inspection regarding the ‘state’ of the wind turbines. Examples of such power curves are shown in Fig. 4, where it can be seen that the decision whether they are ‘good’ or ‘bad’ simply based on visual inspection, is debatable: Fig. 4(a) displays power curtailment, and a few zero power values, whereas Fig. 4(b) displays scatter in the power values. All power curves shown in Figs. 3–4, are normalised to a zero mean value, and a standard deviation of unity, both in active power as in wind speed. In total there were 55 weeks in the ‘normal/good’ class, 36 in the ‘abnormal/bad’ and 34 in the ‘ambiguous/unidentified’ class.

The examples of the ‘bad’ power curves (as those were separated based on information from Vattenfall at the initial stage) shown in Fig. 3, which were correctly identified as ‘bad’, display a lot of zero/negative power values with high wind speed, and this indicates that the turbine was either not working at all, due to severe faults, or was fully shut down, potentially due to maintenance. Since there was no information as to what actually happened, such a week had to be classified as ‘bad’. Although the identification of such events may seem trivial, the proposed method should be able to distinguish between power curves that do not contain zero values and deviate from the normal. The example of a ‘bad’ power curve shown in Fig. 1, which was used as reference, also contained zero values with high wind speed. The data used in this study, are limited to what actually happened to that turbine during the period of data recording, and the majority of them do contain some zero/negative power values. It is also reminded that there is no information regarding maintenance or actual faults for any of the other turbines corresponding to the same wind farm, and consequently, even though other turbine sets may contain different ‘bad’ curves, they are not fully explored, and are not currently used in this study. The value of the proposed

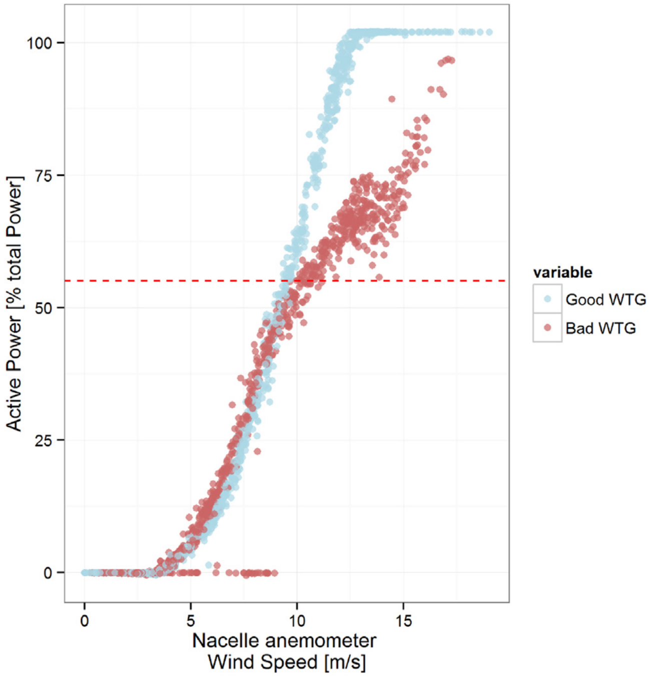


Fig. 1. An example of a ‘good’ and ‘bad’ power curve. WTG stands for wind turbine generator. Figure provided by Vattenfall.

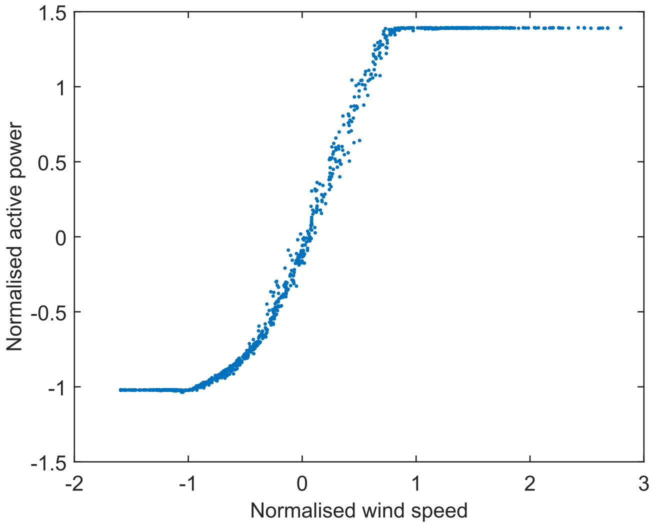


Fig. 2. An example of a power curve considered to belong to the ‘normal’ class, corresponding to a week of data, which was also identified as ‘normal’ by the EFT approach.

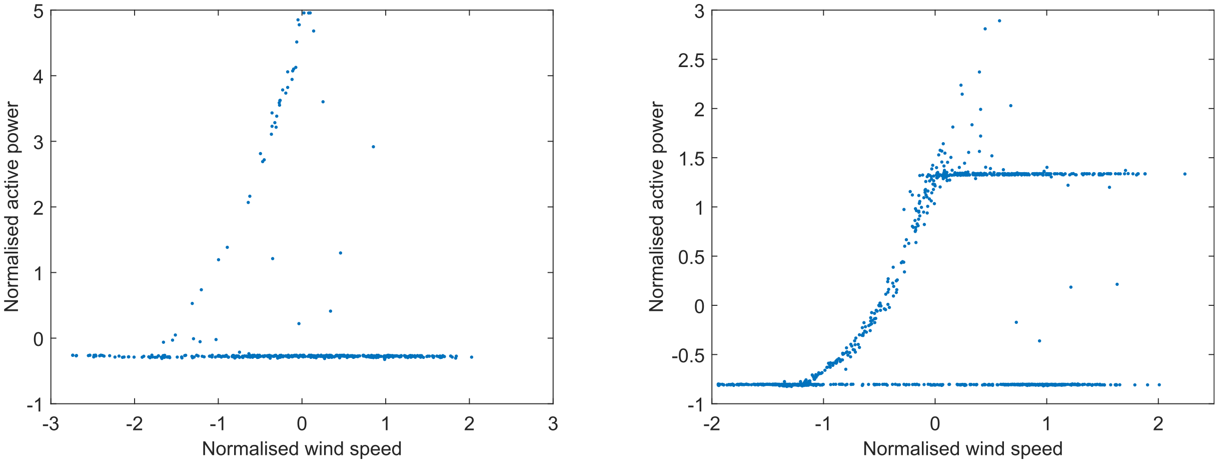


Fig. 3. Examples of power curves considered to belong in the 'abnormal' class, each corresponding to a week of data, which were correctly identified as 'abnormal' by the EFT approach.

approach should be considered in line with the unclassified group, and will be discussed in the final section of this paper.

3.2. General methodology for the classification of power curves with extreme function theory on wind turbine data

The initial step in the methodology after creating the datasets is to train a Gaussian process (GP) based on data from the normal weeks. All the training was executed with a Matlab package provided by Ref. [25]. A squared-exponential covariance function (shown in Equation (10)), and a zero-mean function were the choices for the GP. The covariance function contains some parameters, commonly referred to as hyperparameters (see Equation (11)) which are obtained during the stage of training by minimising the negative log marginal likelihood of the training data. For more details on GPs the reader is referred to [25].

The following step is to use GP regression in order to create sets of z (probability density) values (see Section 2.3) where the differential evolution (DE) algorithm will fit extreme value (EV) distributions - this process is explained in detail in the following subsection (Section 3.3). Having acquired EV parameters, then, it is trivial to obtain thresholds and use GP regression (as described in Section 2.3) on a testing week in order to declare a week as extreme or not.

In order to create a consistent overall approach for the classification of power curves at a weekly resolution, the above steps are summarised here:

- Let $\{n, n_s, a\}$ be the three classes to which the turbine data were separated, where $\{n\}$ is the 'normal' class, $\{a\}$ the 'abnormal' and $\{n_s\}$ the non-categorised class, and $\mathbf{p_i} = \{\mathbf{x_i}, \mathbf{y_i}, i = 1...N\}$ is a sampled power curve (corresponding to a week of data in this work). Then a training set $Sp_i = \mathbf{q_i} = \{\mathbf{x}'_\mathbf{i}, \mathbf{y}'_\mathbf{i}, i = 1...N_t\}$ can be created where the $\mathbf{x}'_\mathbf{i}$ are sampled randomly from the $\mathbf{x_i}$. As $N_t > N$, a bootstrap approach is effectively applied.
- Create three sets: training $T_r = \{\mathbf{q_i}, \mathbf{q_i} \subset \{n\}\}$ sampled randomly from $\{n\}$, of size $N_c \times N_s$, where N_c is the number of training power curves (weeks) and $N_s = N_t$, a validation set $V = \{\mathbf{q_i}, \mathbf{q_i} \subset \{n\}\}$ such that $V \cap T_r = \varnothing$, and a testing set $T_e = \{\mathbf{q_i}, \mathbf{q_i} \subset (\{n\} \cup \{n_s\} \cup \{a\})\}$, making sure that it is different from the training and the validation sets.
- Train a Gaussian Process (GP) entirely on T_r and use the V set to create sets of z values where the DE algorithm will fit extreme value (EV) distributions. Choose the best EV parameters based

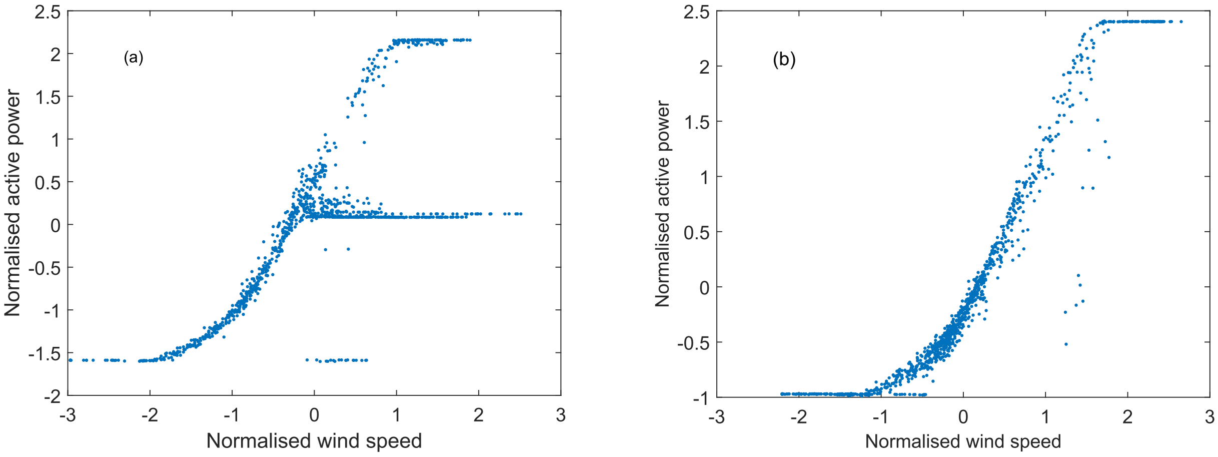


Fig. 4. Examples of power curves, corresponding to a week of data, which were originally considered unidentified (class n_s) and later classified as 'abnormal' (a), and 'normal' (b) by the EFT approach. Both were classified as 'abnormal' by the conventional pointwise approach.

on the lowest normalised mean squared error (NMSE), and then test on T_e (with the thresholds created from the selected EV distribution) for the classification of the power curves (at a weekly resolution). Alternatively, a separate validation set V' may be created, sampled from the $\{n\}$ and $\{a\}$ classes to be used for the selection of the EV parameters. In that case, the choice of EV parameters can be simply based on the best classification rate on V'.

3.3. Fitting EV distributions with DE

As differential evolution (DE) belongs to the family of evolutionary algorithms, several runs may be needed to reach an optimal, or near optimal, solution. In the turbine data used in this study, there were in total 55 weeks identified (using Fig. 1 as a reference) as ‘normal’, 36 as ‘abnormal’, and the 34 rest as the unidentified (class $\{n_s\}$). In order to have sufficient data to obtain accurate results, a bootstrap methodology was applied here, but only in the creation of the validation set, not the training. This means that there was a repeating process of randomly subsampling from the validation set. The number of sample points from each week was kept to 50, mainly because larger numbers could cause numerical issues, but also because it is computationally and practically attractive to perform classification based on data of small size. The choice of size of the sets can be arbitrary, but one should consider having sufficient amount of data for training, and eventually for testing. Here, from the 55 normal weeks, 6 were used for training (corresponding to 300 sample points from the curves), and 35 to create the validation set V, which was then used for the estimation of the parameters of the extreme value (EV) distribution based on the best fit. Because the values of z were very small, the optimisation was performed based on its natural logarithm, $ln(z)$. By using this quantity, there is no guarantee that its convergence (after infinite samples) will be the Weibull DF (as was shown in Ref. [24]), so all three EV CDFs (Gumbel, Weibull, Frechet) were investigated. An alternative would be to use the generalised EV distribution [32]. After the algorithm was initially left to run, the best results were obtained with a Gumbel distribution, which is given by,

$$H(x) = exp^{-exp^{-\frac{x-\lambda}{\delta}}} \tag{17}$$

$$L(x) = 1 - exp^{-exp^{\frac{x-\lambda}{\delta}}} \tag{18}$$

where $H(x)$ is the distribution used for the maxima, and $L(x)$ for the minima. Because extreme functions need to be identified here, only the $L(x)$ was employed (i.e. extreme events that have a lower probability density).

Figs. 5 and 6 show examples of a Gumbel distribution curve-fit after the DE was left to run and fit to different sizes of data sets from the parent distribution. It can be seen that there can be a difference in the curve-fits with the size of the portion of the data which is used for fitting - something expected. In addition, variability on the curve-fit and the NMSE values was also observed when the process was repeated with different seeds of random samples (it is reminded that all sets are created randomly). It was then decided to make use of the second validation set V' in order to increase the consistency of the selection of the EV parameters. The size of the portion of the data from the V set which was used for fitting was set at 10%, and the size of the second validation set V', was 30 weeks. The objective function of the DE algorithm now was simply the ratio of the classification rate (based on V') divided by the NMSE. After several runs the algorithm settled on the EV fit shown in Fig. 6, which is a nice result indicating that the original fitting was consistent. With the parameters of the Gumbel distribution identified, a threshold was calculated for a confidence interval of 99%, meaning that 1% of the normal functions will be considered outliers. Fig. 7 shows the results of this novelty detection scheme on the set of the test weeks. It is reminded that the test set contains 9 normal weeks which were not used neither in the training nor in the estimation, the group of weeks which could not be identified, and finally the abnormal as last. For the purposes of illustration the negative logarithm of the density z is shown. It can be seen that there is no misclassification in the 9 normal weeks, and the majority of the unidentified weeks are classified as normal. From the abnormal weeks, 4 (out of 36) were incorrectly identified as normal. Fig. 3 actually shows an example of a ‘bad’ week which was correctly identified as abnormal.

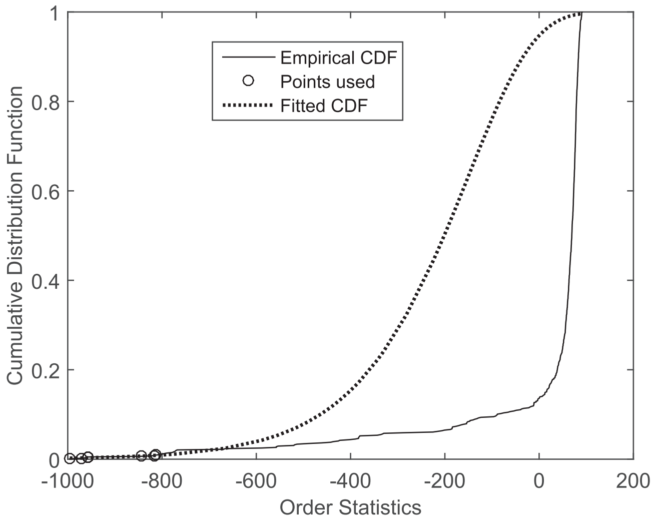


Fig. 5. An example of a Gumbel distribution curve-fit after the Differential Evolution algorithm was left to run, fitting on 1% of data.

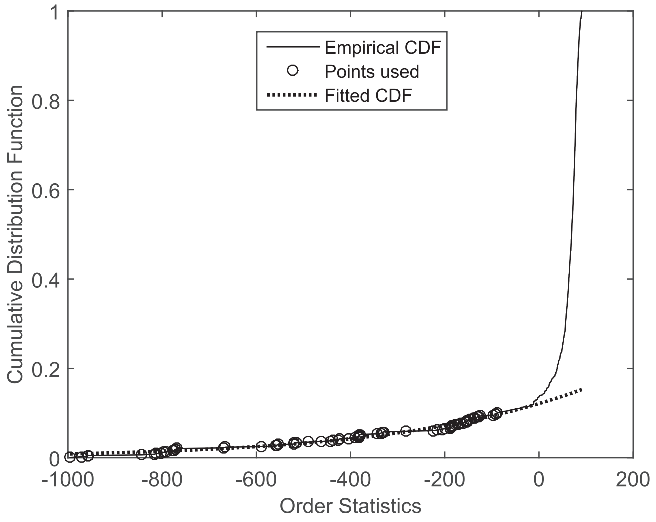


Fig. 6. An example of a Gumbel distribution curve-fit after the Differential Evolution algorithm was left to run, fitting on 10% of data.

3.3.1. Effect of random sampling on the wind turbine data

Since it was shown that sampling randomly (from the ‘normal’ $\{n\}$ and ‘abnormal’ $\{a\}$ class datasets) to create the training, validation and testing sets may have an effect on the fitting of the EV

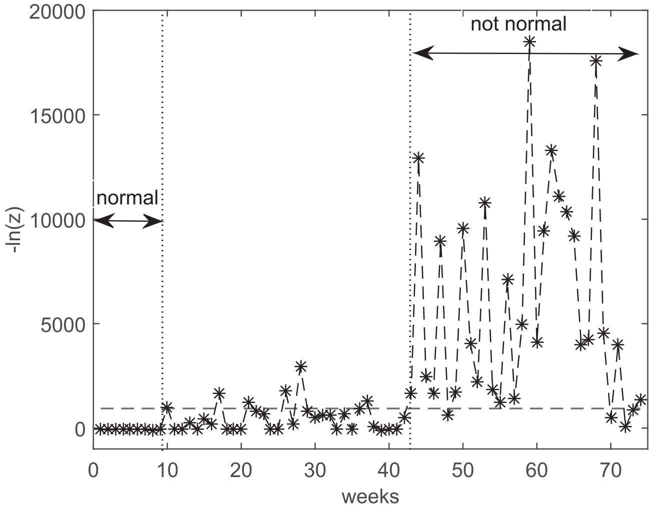
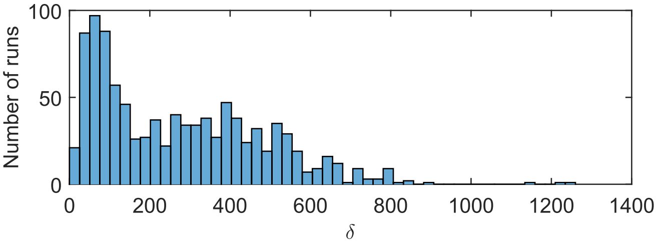


Fig. 7. Novelty detection at a weekly resolution on wind turbine data with extreme function theory. The curves are ordered with weeks corresponding to the 'normal' set first, and 'abnormal' set last.



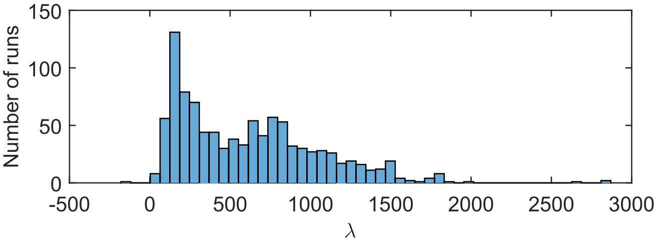


Fig. 9. Histogram of the parameters of the fitted Gumbel distributions after 1000 runs of different sampling from the turbine data.

CDFs, it was decided to perform a short analysis to investigate that effect. The process of randomly creating the three sets $\{T_r, V, T_e\}$ was repeated 1000 times and the DE algorithm was run each time in order to fit Gumbel CDFs. The second validation set was not used, as this was a process to see the effect of the random sampling in the parameters of the EV distributions. Fig. 8 shows a histogram of the NMSE values of the best results from all the 1000 runs, and Fig. 9 shows a histogram of the actual parameters that were identified. It can be seen that the NMSE values vary, but the majority of them are between 1 and 5, it is noted here that an NMSE value below 5 indicates a good fit, and below 1 an excellent fit [33]. Although there are cases with bad fitting, there are also cases with excellent results.

Fig. 10 also shows a 3D plot of the identified parameters with the NMSE values, and it can serve as a way to see where the parameters should lie in order to have low NMSE values. It is also obvious that the parameters may vary significantly, and that is probably the effect of the nature of the data used, the logarithm values of the z quantity given by Equation (13). Nevertheless, Fig. 10 can be used to select constraint values for the identified parameters, especially when an evolutionary algorithm like DE is used; for example there is a higher concentration of lower NMSE values at the area where λ lies between 20 and 200, and δ between 0 and 500. The whole process was repeated with the DE algorithm fitting Weibull distributions, and the results were equivalent to those presented here (with the Gumbel) so it was not considered necessary to display those results.

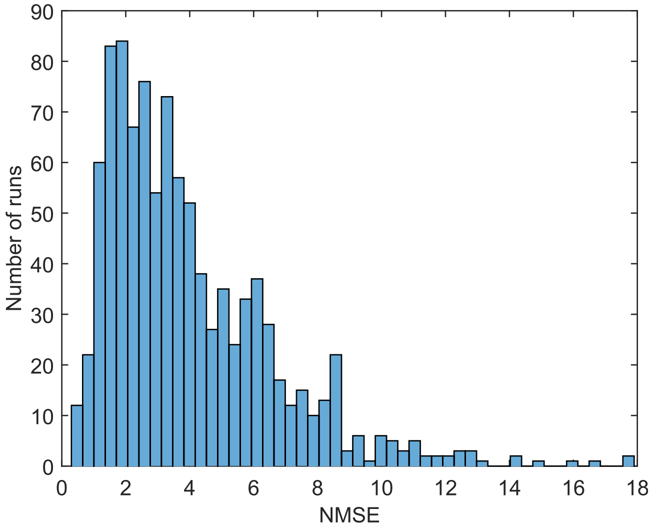


Fig. 8. Histogram of NMSE values after 1000 runs of different sampling from the turbine data.

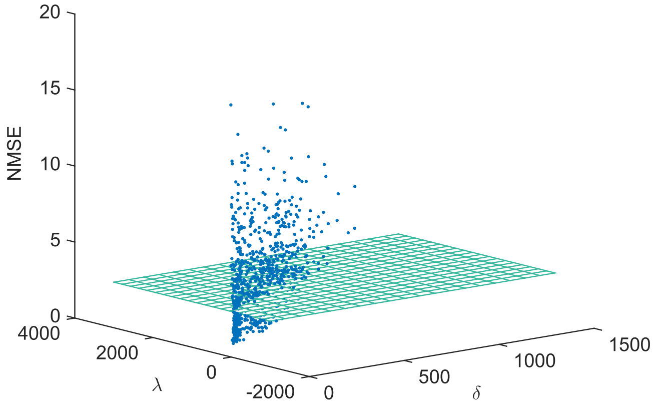


Fig. 10. A 3D plot of the parameters of the fitted Gumbel distributions versus the NMSE after 1000 runs of different sampling from the turbine data. The surface slice corresponds to NMSE of 3.

4. Comparison with other conventional approaches

4.1. Conventional pointwise approach

The novelty detection results shown in Fig. 7 seem very promising, as the method was applied to actual data from wind turbines with a good identification ratio. In order to properly assess the method, and the idea of classifying functions instead of data points, a comparison with a conventional pointwise approach was deemed necessary. As mentioned earlier, most such approaches require a quantity to be compared against a threshold, usually in combination with a statistical test. A very common such quantity is, or is derived by, the residuals between the 'normal' model predictions and the actual data. This standard approach was followed here as well, where Gaussian Processes were again the chosen algorithm for modelling of the normal power curves. The residuals of the GP predictions and the data from the weeks were monitored in a similar approach as in Ref. [19] where control charts had been applied. The critical values (threshold) for the monitoring of the residuals were calculated exactly as in Ref. [19] with the help of EV statistics. The overall approach can be summarised as follows:

- Create a training set $T_r = \{\mathbf{q_i}, \mathbf{q_i} \subset \{n\}\}$ sampled randomly from the 'normal' set $\{n\}$, and train a GP.
- Use the residuals of the training set to fit EV distributions with the help of the DE algorithm, and get critical values (upper and lower control limits) for a desired level of confidence.
- Feed the 'normal' model with data from all the weeks and monitor the residuals. Declare a week as not normal when the number of outliers exceeds the level of confidence decided in the previous step. The use of a testing set T_e sampled from all weeks, but different than the training set is recommended for the proper assessment of the approach.

For the sake of comparison with the EFT approach here, the training set used was identical to the one used in the EFT of the earlier section (see Fig. 7). Fig. 11 displays a bar plot with the percentage of outliers produced when data from all the 125 available weeks were fed to the fitted GP. As said above, the critical values were chosen with the use of differential evolution and EV statistics, and the Gumbel distribution was the choice again as it presented the lower NMSE values in the fitting. The level of confidence chosen was 99%, meaning that 1% of normal data are expected to go outside the control limits. Fig. 11 also shows the 1% value for the percentage of outliers, and it is clear that there are very few weeks which present less than 1% of outliers, in fact there are only 15, something which indicates a high number of false-positives. Of course, a fair comparison with the EFT approach presented in the previous section should make use of the exact testing set T_e chosen there, so Fig. 12 displays exactly that. It can be seen that in that testing set there are only 4 weeks which contain less than 1% of outliers, and therefore are declared as 'normal'. From the 9 weeks sampled from the 'normal' (class $\{n\}$) set, only 3 are correctly classified as 'normal' which makes the number of false-positives (FP) very high.

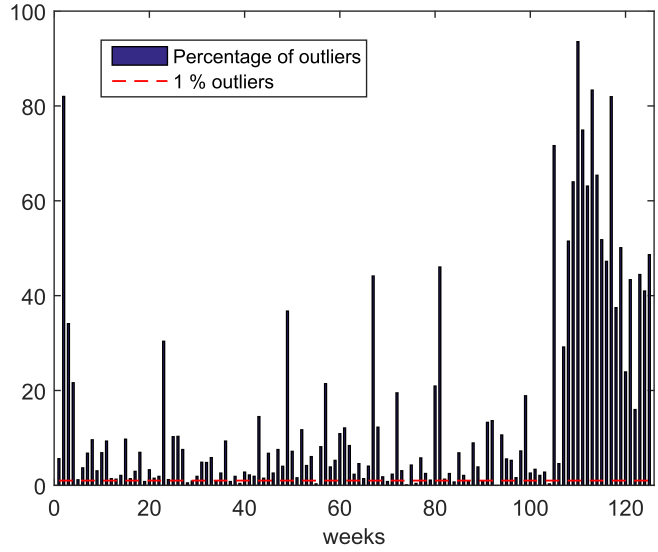


Fig. 11. Bar plot displaying the percentage of outliers (threshold crossings) in all 125 weeks from the conventional pointwise approach.

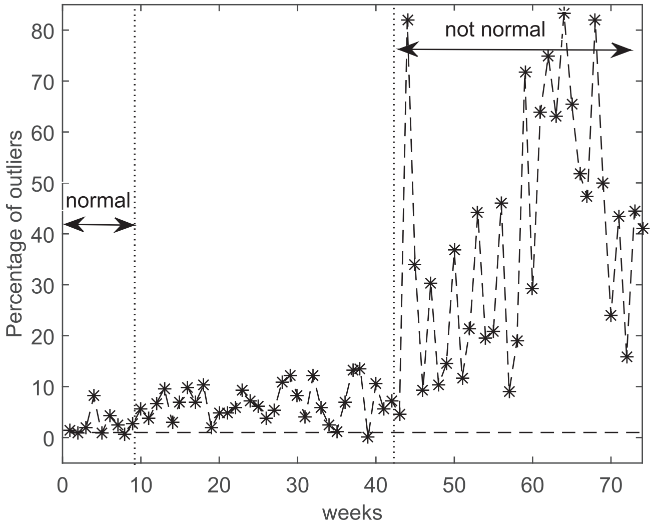


Fig. 12. Percentage of outliers in the testing set used in the EFT approach with the conventional pointwise GP methodology. Critical value is set at 1%, and the weeks are ordered with the 'normal' first and the 'abnormal' last.

4.2. Comparison with approaches using multivariate features

The EFT approach has been shown to significantly decrease the false-positive (FP) rate when compared to a more conventional pointwise approach where single values were used. It is also compared here with a more conventional approach that makes use of multivariate features instead of just single datapoints. For such an attempt, GPs are again the chosen method of modelling the 'normal' power curves. In order to create multivariate features the following steps were applied:

- Create a training set by randomly sampling n_d points from a 'normal' week. Train a GP and then choose d equally spaced wind speed points (between the minimum and maximum actual wind speeds) and get the power predictions from the GP.
- Repeat the process m times, and obtain another set of power predictions for the exact same wind speeds as above.
- Optionally, repeat the process for another 'normal' week. There is now a multivariate feature containing predictions of power for the exact same wind speeds of size $b \times n$, where b is the number of selected 'normal' weeks multiplied by m, which corresponds to the normal state of the wind turbine.
- Repeat the process of training GPs and obtain power predictions for each of the 125 weeks. There is now a testing set of size $125 \times d$.

4.2.1. Outlier analysis

By following the above procedure, multivariate features of the 'normal' state can be created, and a test set for all weeks can be assessed in terms of novelty. There could be various ways of assessing the test set, but here outlier analysis [23] was the chosen method. Every multivariate feature can be fused into a single quantity called the Mahalanobis squared-distance and subsequently compared against a critical value (threshold) to be declared as an outlier or not. The Mahalanobis squared-distance is given by,

$$D_\zeta = \left(\mathbf{x}_\zeta - \overline{\mathbf{x}}\right)^{\mathbf{T}} \mathbf{S}^{-1} \left(\mathbf{x}_\zeta - \overline{\mathbf{x}}\right) \qquad (19)$$

where $\mathbf{x}_\zeta$ is the potential outlier, $\overline{\mathbf{x}}$ is the mean vector of the sample observations, and $\mathbf{S}$ the sample covariance matrix. The threshold depends both on the dimension of the problem and on the number of observations, and it can be exclusive or inclusive depending upon whether the sample being tested was included or not in the computation of the sample statistics (Mahalanobis squared-distance here).

4.2.2. Monte Carlo threshold

The calculation of the threshold, for a multivariate case here, is based on a Monte Carlo procedure which was followed for its calculation exactly as in Ref. [23].

In order to compare directly this approach with the EFT, the number of the random sample points n_d used for the training of each week was kept at 50 (as was done in the EFT approach). The number of the dimensions of the features d varied between 10 and 30, since it was found that above 30 there were numerical issues in the calculation of the Mahalanobis distance. For the creation of the normal feature, the process of training on the same week was repeated 10 times. Fig. 13 displays the Mahalanobis distances for all 125 weeks when the multivariate features were created with $n_d = 50, m = 10$, and $d = 10$. With those choices the normal feature had a size of 60 × 10, since the training set used was identical to the one used for the EFT approach (and contained 6 ‘normal’ weeks). The exclusive threshold in this case was calculated to be 52. It can be seen that there is a higher rate of false-negatives at 9 out of 36 when compared to the EFT approach (4/31), and more false-positives (6/55) (see Fig. 13).

When the exact testing set which was used in the EFT method is also applied here, the number of false-positives dropped to one (out of 9), see Fig. 14. As was mentioned earlier, the values of the dimension d tried were kept between 10 and 30, but they did not produce any significant change to the classification results and so they are not shown. The increase of the sampling points n_d for the training of the GPs may improve the number of the false-negatives slightly. It should be noted that with this approach it is necessary to train GPs for each week, and several times for the weeks of the training set, whereas in the EFT approach the training is only done once.

4.2.3. Calculating an EV threshold with DE for outlier analysis

The second approach for calculating a critical value for the Mahalanobis squared-distance made use of the DE algorithm. In a similar way as in the EFT approach shown in the previous section, the DE algorithm was used to fit EV distributions in the multivariate feature representing the normal state of the turbines (normal weeks). Again, all three EV distributions, Gumbel, Weibul and Frechet were investigated, with the Gumbel giving the best results. In order to have more values for fitting, the 20% higher values were used (compared to the 10% used throughout the rest of the work). Fig. 15 shows the Mahalanobis squared-distance for the same testing set (weeks) as in the EFT approach. It can be seen that the threshold is slightly lower than in the Monte Carlo approach (42 compared to 52), with no changes in the FP/FN rate.

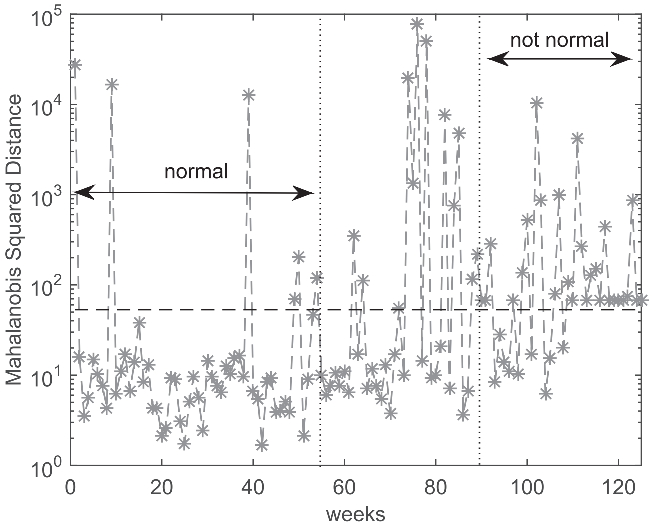


Fig. 13. Outlier statistics for each of the 125 weeks after using multivariate features created through GP regression. Monte Carlo threshold.

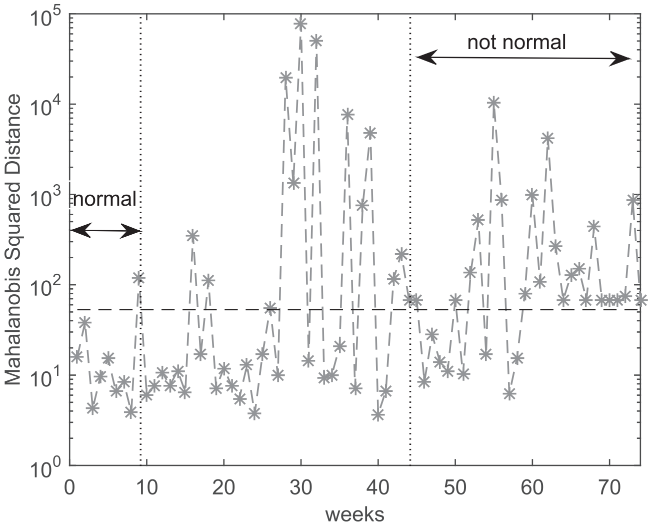


Fig. 14. Outlier statistics for the same testing set (weeks) used in the EFT approach with a Monte Carlo threshold. Weeks are ordered with the ‘normal’ first and the ‘abnormal’ last.

4.2.4. Auto-associative neural networks

The outlier method implicitly assumes that normal condition data are Gaussian. This can be a source of error, so it is useful to compare with a more general approach. A general method that can be applied in advanced unsupervised novelty detection analysis is the auto-associative neural network (AANN) [2,34–36]. This neural network architecture was motivated by Nonlinear Principal Component Analysis (NLPCA) which is a robust and powerful statistical method for feature extraction and dimension reduction. The AANN is a type of multi-layer perceptron MLP whose target outputs are the same as the input. Generally, the AANN consists of five layers including the input, mapping, bottleneck, demapping and output layers [35–40]. A restriction of the mentioned topology is that the bottleneck layer must have less neurons than the input and output layers, and this performs compression as the AANN must reproduce the input vector at the output.

When a trained AANN is given an input feature vector set coming from an unprecedented state of the structure, a novelty index n_i described in the form of Euclidean distance will increase,

$$n_i(\mathbf{y}) = ||\mathbf{y} - \hat{\mathbf{y}}|| \quad (20)$$

where $\mathbf{y}$ and $\hat{\mathbf{y}}$ are each initial input and network output vectors respectively. If the neural network learning was successful then $n_i(\mathbf{y}) \approx 0$ for all the training data set. Later on in testing, $n_i(\mathbf{y})$ may significantly depart from zero indicating the presence of novelty i.e. $n_i(\mathbf{y}) \neq 0$. The warning levels can be defined as $\overline{n}_i + a\sigma$ where $\overline{n}_i$ and σ are, respectively the mean and standard deviation of all the values of the novelty index over the training data. In statistical terms, the

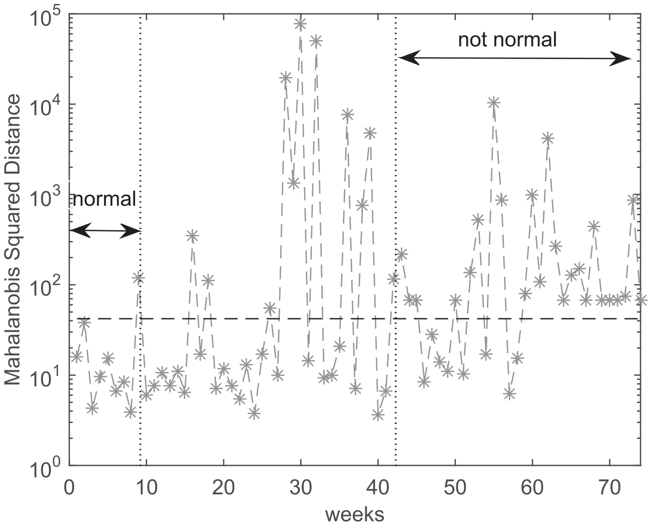


Fig. 15. Outlier statistics for the same testing set (weeks) used in the EFT approach by using a threshold calculated with DE. Weeks are ordered with the 'normal' first and the 'abnormal' last.

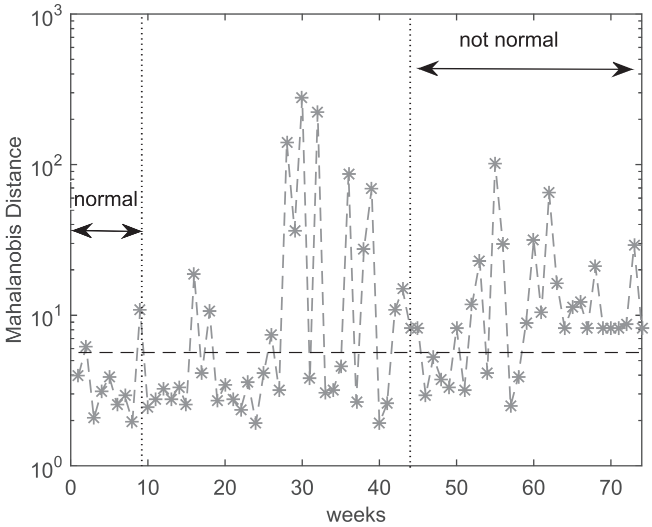


Fig. 17. Applying the multivariate extreme value threshold on the Mahalanobis distance on the same testing set as in the EFT approach. Weeks are ordered with the 'normal' first and the 'abnormal' last.

parameter a controls the percentage of false-positives. For example, if the distribution is purely Gaussian, then a value of 1.96 will give estimates within warning levels of 95% probability. In this paper, a is set equal to 2.58 giving a 99% confidence limit. Fig. 16 shows the Euclidean distance between the output of the auto-associative neural network and its target for the same testing set as in the EFT approach. There is one FP (out of 9), and 5/31 FN.

4.2.5. Multivariate extreme value theory

Clifton et al. [41] have presented the use of extreme value theory (EVT) for multivariate cases. This approach can also be applied here in order to fit an extreme value distribution from the multivariate features, and obtain critical values for arbitrary levels of confidence. The multivariate features are assessed in terms of their Mahalanobis distance, and the approach is related to the exact extreme function theory (EFT) shown in Ref. [24], with the difference that multivariate features are used, and not functions. Fig. 17 shows the Mahalanobis distance for the same testing weeks as in the EFT approach presented in the previous section, there are 2 out of 9 false-positives, and 8 out of 31 false-negatives which is higher than the EFT case.

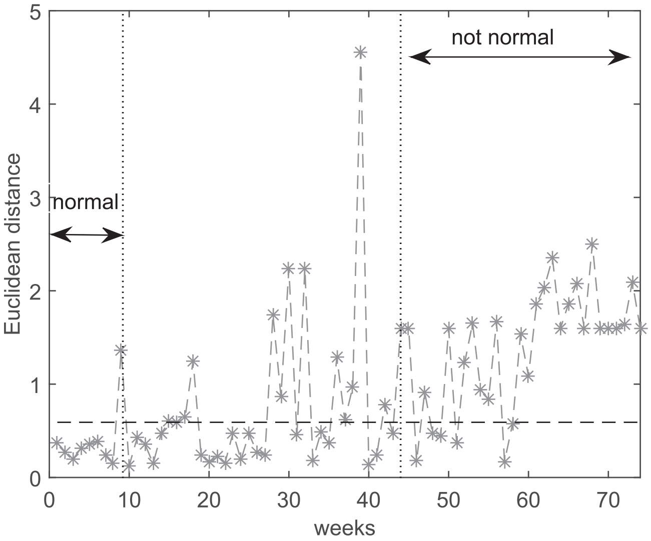


Fig. 16. Euclidean distance between the output of auto-associative neural network and its input (target) for the same testing weeks as in the EFT approach.

Table 1 shows the overall classification results of all the approaches that were compared. The FP rate corresponds to the number of FP divided by the total number of normal weeks assessed in the testing set, in this case 9, and the FN rate is the number of FN divided by 31 (total number of 'abnormal' weeks in the testing set). Classification error rate is simply the sum of FP and FN divided by the total number of weeks assessed in the testing set (here 40). Sensitivity or true positive rate (TPR) is the number of true-positives (TP) divided by the total number of 'abnormal' weeks. TP are the weeks which were correctly classified as 'abnormal', hence sensitivity is a measure of how 'sensitive' is the approach to damage (or 'abnormality' in general). Specificity (SPC) or true negative rate is the number of true-negatives (TN) divided by the total number of 'normal' weeks. TN are the correctly identified 'normal' weeks, so specificity is a measure of how likely the method is to misclassify 'normal' weeks. Ideally, one would prefer a method that performs with 100% in both specificity and sensitivity. The exact formulas for all the error measures are given below,

$$FP\ rate = \frac{FP}{FP + TN} \tag{21}$$

$$FN\ rate = \frac{FN}{FN + TP} \tag{22}$$

$$CE\ rate = \frac{FP + FN}{TP + FP + FN + FP} \tag{23}$$

$$sensitivity\ (TPR) = \frac{TP}{TP + FN} \tag{24}$$

$$specificity\ (SPC) = \frac{TN}{TN + FP} \tag{25}$$

It is clear from Table 1 that the EFT approach is superior in the overall classification error (10% when the rest are 15% and above), but most importantly in the low number of FPs and the high

Table 1
Classification error (CE) rates for all the approaches.

	EFT	Pointwise GP	GP with Monte Carlo	GP with DE threshold	Multivariate EVT	AANN
FP rate	0	0.67	0.11	0.11	0.22	0.11
FN rate	0.13	0	0.26	0.26	0.26	0.16
CE rate	0.1	0.15	0.23	0.23	0.25	0.15
TPR	0.87	1	0.74	0.74	0.74	0.84
SPC	1	0.33	0.89	0.89	0.78	0.89

number of both the sensitivity and specificity. As was seen in Fig. 12 the conventional pointwise approach is 100% sensitive, but has a lot of FPs with a specificity of 33%.

Fig. 18 shows the use of PCA analysis for the visualisation of the full multivariate test feature. The whole feature (corresponding to all 125 weeks) is projected on its two first principal components in order to assess how well each class ('normal', 'abnormal' and 'unclassified') separates from each other. It can be seen that many of the 'unclassified' weeks overlap with the 'normal', but some overlap with the 'abnormal'. Also some 'abnormal' weeks overlap with the 'normal', and this can provide a possible explanation for all the results shown where it seemed that reducing the FP rate would increase the FN rate. However, the EFT approach was shown to reduce the FP by not sacrificing so much as the other approaches in the number of FN.

It was mentioned earlier that sampling randomly for the creation of the sets may have an effect on the approaches, so all the process of applying the EFT approaches, the pointwise, and the multivariate features was repeated 50 times. Table 2 shows the mean values of the classification error rates for all those 50 repetitions, and for all the methods. It can be seen that the EFT approach displays approximately the same classification error (0.5% lower) as the conventional pointwise approach, but has the lowest FP ratio as can be seen in the 99% specificity value, where the pointwise GP performs with only 43%. It is clear that the EFT reduces the FP to almost zero without sacrificing the sensitivity (84%).

5. Discussion

The FP (false-positive) rates are higher than the test run shown in the previous section, but it should be said that the comparison is made on exactly the same testing set always, and with the same settings (number of training, validation and second validation curves) as in the case shown in Section 3.3. The approaches which use multivariate features show more diversity in their results with the repetition of the test than the EFT as can be clearly seen in the histograms of Fig. 19 for the number of FP (Table 2 only shows mean values). The increase of the sampling points from each week in the creation of the multivariate features will improve slightly those results, and it will reduce the FN rate slightly, but it was kept at 50 to maintain a fair comparison with the EFT. The process of creating multivariate features demands the training of GPs several times, and for each week, whereas with the EFT it is done only once. It should also be mentioned that the process of training the AANN requires careful consideration, and a lot of input from the user, something which was not done during the 50 test runs, so some of the results of the AANN could probably be improved - on average the AANN is better than multivariate EVT. At the same time, the same could be said for the EFT approach, as when the DE algorithm was run, its settings were kept constant, meaning that there is no absolute guarantee an optimal solution was always reached. Overall, the confidence level used in the EFT corresponds to a function, so in this case a 'week', and not a single datapoint corresponding to a 'week' as in a pointwise approach. The EFT approach is ideal for a realistic problem of monitoring wind turbines at a weekly resolution.

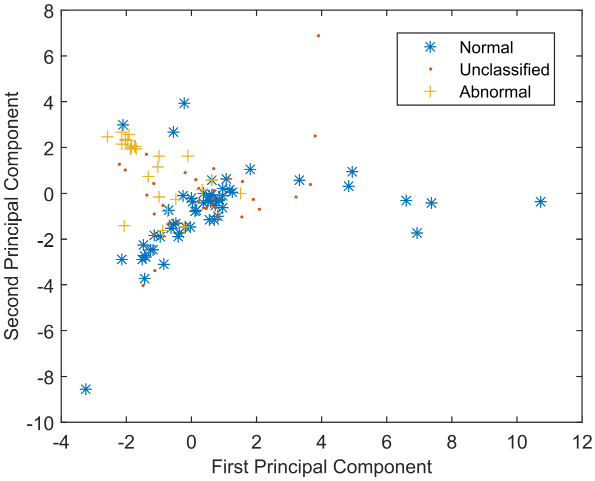


Fig. 18. Projection of the multivariate test feature on its first two principal components.

A very important point here is the treatment of the unidentified class $\{n_s\}$ of weeks. These were power curves which were not easily identified by visual inspection as either 'normal' or 'abnormal', and the need for objectivity would be paramount in any such real attempt at monitoring. Fig. 4 shows two examples of such curves, and it can certainly be debated whether they belong to the 'normal' or 'abnormal' class. It can be seen from Fig. 7 that the EFT approach classifies most of those weeks as 'normal' with 6 out of 34 identified as 'abnormal'. It is clear from Fig. 12 that the conventional GP pointwise approach classifies almost all (32 out of 34) as 'abnormal' meaning that it is very sensitive to a deviation from the normal model, something shown in the high FP (false-positive) rate. Such a method would not be ideal for monitoring real wind turbines, as the high number of false alarms would increase the maintenance costs and obviate the benefits of monitoring.

The original separation of the weeks into classes did contain a degree of subjectivity, although reference curves were used, but it should be said that any data originating from real structures under real, and not laboratory conditions, are expected to have a certain degree of variability which could challenge any novelty detection scheme, and the extreme function theory shows enough promise in dealing with that variability, and in the presence of ambiguous power curves. It is also important that the functions assessed can be represented by samples of different size which would be of value in practice.

The effect of the random sampling in the creation of the training, testing, validation sets (see Section 3.3) may influence the identification of the parameters of the Extreme Value (EV) distribution. However, as the whole process of using and comparing the EFT approach was repeated 50 times (with different training/testing sets randomly created each time), and was shown to be robust (see Table 2), it can be reasonably argued that the uncertainty due to

Table 2
Mean classification error (CE) rates for all the approaches after 50 tests.

	EFT	Pointwise GP	GP with Monte Carlo	GP with DE threshold	Multivariate EVT	AANN
FP rate	0.01	0.57	0.16	0.14	0.23	0.29
FN rate	0.16	0	0.36	0.44	0.22	0.13
CE rate	0.125	0.13	0.32	0.37	0.23	0.17
(TPR)	0.84	0.99	0.64	0.56	0.78	0.86
(SPC)	0.99	0.43	0.84	0.86	0.77	0.71

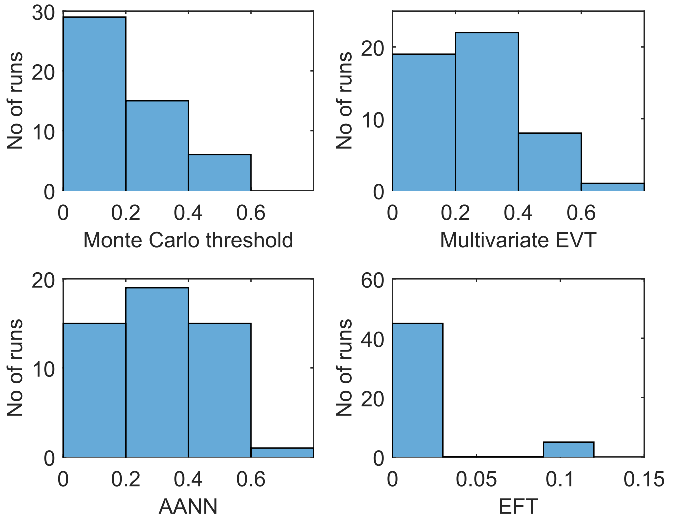


Fig. 19. Histogram of the FP rate of some of the approaches after 50 tests.

random sampling is not important. The data used were real from a period of 125 weeks including data from all seasons, so uncertainty due to environmental effects or noise through the SCADA data did not play an important factor. The sensitivity of the approach depends entirely on the sensitivity of the power curve monitoring - as it was used as an established methodology here. Ultimately, all the approaches compared here were comparing a ‘normal’ power curve from an ‘abnormal’, and it was shown that the EFT is ideal in minimising false alarms without sacrificing a lot its sensitivity.

6. Conclusions

This paper presented a modified version of the extreme function theory (EFT) for the performance monitoring of a wind turbine at a weekly resolution over a period of 125 weeks. The idea was to test individual functions in terms of normality, and not just individual points as would happen in a pointwise approach. The functions were simply the power curves (power versus wind speed) from weekly data. In this way a model of normality, represented here by a Gaussian process (GP), was built based on weeks which were considered normal, and then a novelty detection scheme was successfully applied to a testing set. The main idea is to test how likely a test function is to originate from the GP representing the normal condition, and how extreme it is compared to it. The testing set obtained normal, abnormal and unidentified weeks (power curves). In the main case presented here, there were no false-positives, meaning ‘normal’ weeks which are wrongly identified as ‘abnormal’, and the majority of the unidentified weeks were considered normal; there were also 4 out of 31 false-negatives, meaning weeks which were incorrectly classified as normal. The method was compared to a conventional pointwise approach which made use of GPs again and exploited the residuals created from the model of normality when compared to the data from the turbines in a weekly resolution, and clearly showed that it is superior in terms of the false-positive rate. Another comparison made use of multivariate features which were created again with the help of GPs and were assessed in four different ways with the help of outlier analysis, extreme value statistics and an auto-associative ANN, and confirmed that the extreme function theory displays a low false-positive ratio, and this can be attributed partly to the idea of classifying functions (represented by time series data sets) and not just individual points. The comparison of all the approaches was repeated for 50 experiments, and showed that the classification error of the extreme function theory was on average low, although only marginally lower than the pointwise approach. However, the small number of normal weeks used in the testing set, when compared to the number of abnormal weeks used, affects the overall classification error rate. The number of false-positives was also very low, almost zero, for the extreme function theory which can be translated to an excellent specificity, meaning how likely the method is to misclassify normal curves, without sacrificing a lot of sensitivity, as would happen with a simple raise of a threshold in the conventional pointwise GP (equivalent to a control chart) approach. It should be noted that the conventional pointwise method used here already makes use of extreme value thresholds, which tend to be higher than standard Gaussian thresholds.

Equally important is the unidentified class of weeks, which were mostly identified as normal with the EFT approach, but ‘abnormal’ with the conventional pointwise GP methodology. The unidentified weekly power curves can be arguably considered to lie at the core of the importance of this work, since there is a need for objectivity in their classification - otherwise a visual inspection would be considered enough. Without any information regarding actual faults on the wind turbines, the unidentified weeks cannot be unambiguously confirmed as ‘normal’ or ‘abnormal’, although it is possible that this would be challenging even with such information available, so one would have to use the ‘normal’ and ‘abnormal’ testing sets to validate the approach, and subsequently choose to either trust or discard the decision on the unidentified class. Based on the high FP rate of the conventional pointwise method, and the fact that almost all the entirety of the unidentified class was considered ‘abnormal’ by said method, it does not seem the best for the classification of the ambiguous weeks, whereas the EFT approach seems better suited for that task.

Finally, the work presented here, does not address neither the suitability nor the sensitivity of the power curve method, it is applying it as an established monitoring approach, and in combination with the extreme function theory for the first time. Such a performance monitoring approach may not be suitable to identify sub-critical fatigue cracks on turbine blades, but can be nevertheless of practical value to the maintenance of a real wind farm, and as it was shown, the extreme function theory can be very well suited for the monitoring of turbines at a weekly resolution. In the future, the theory can be extended to incorporate functions of disparate sources, such as data acquired under different environmental conditions, into the model of normality, and be of even more practical value.

Acknowledgment

The support of the UK Engineering and Physical Sciences Research Council (EPSRC) through grant reference numbers EP/J016942/1 and EP/K003836/2 is gratefully acknowledged.

References

[1] J.E. Doherty, Handbook on Experimental Mechanics, Society for Experimental Mechanics, Inc., 1987. Ch. 12, Nondestructive Evaluation.

[2] N. Dervilis, M. Choi, S. Taylor, R. Barthorpe, G. Park, C. Farrar, K. Worden, On damage diagnosis for a wind turbine blade using pattern recognition, J. Sound Vib. 333 (6) (2014) 1833–1850.

[3] W. Yang, Z. Lang, S. Tian, Condition monitoring and damage location of wind turbine blades by frequency response transmissibility analysis, IEEE Trans. Ind. Electron. PP (99) (2015), http://dx.doi.org/10.1109/TIE.2015.2418738, 1–1.

[4] J. Tang, S. Soua, C. Mares, T.-H. Gan, An experimental study of acoustic emission methodology for in service condition monitoring of wind turbine blades, Renew. Energy 99 (2016) 170–179, http://dx.doi.org/10.1016/j.renene.2016.06.048. http://www.sciencedirect.com/science/article/pii/S0960148116305729.

[5] I. Antoniadou, Accounting for Nonstationarity in the Condition Monitoring of Wind Turbine Gearboxes, Ph.D. thesis, 2013.

[6] W.J. Staszewski, K. Worden, Classification of faults in gearboxes - pre-processing algorithms and neural networks, Neural Comput. Appl. 5 (3) (1997) 160–183.

[7] Y. Lei, J. Lin, Z. He, M.J. Zuo, A review on empirical mode decomposition in fault diagnosis of rotating machinery, Mech. Syst. Signal Process. 35 (1) (2013) 108–126.

[8] D. Yang, H. Li, Y. Hu, J. Zhao, H. Xiao, Y. Lan, Vibration condition monitoring system for wind turbine bearings based on noise suppression with multi-point data fusion, Renew. Energy 92 (2016) 104–116, http://dx.doi.org/10.1016/j.renene.2016.01.099. http://www.sciencedirect.com/science/article/pii/S0960148116300994.

[9] Z. Hameed, Y. Hong, Y. Cho, S. Ahn, C. Song, Condition monitoring and fault detection of wind turbines and related algorithms: a review, Renew. Sustain. Energy Rev. 13 (1) (2009) 1–39.

[10] F.P. García Márquez, A.M. Tobias, J.M. Pinar Pérez, M. Papaelias, Condition monitoring of wind turbines: techniques and methods, Renew. Energy 46 (2012) 169–178.

[11] W. Yang, R. Court, J. Jiang, Wind turbine condition monitoring by the approach of SCADA data analysis, Renew. Energy 53 (2013) 365–376, http://dx.doi.org/10.1016/j.renene.2012.11.030. http://www.sciencedirect.com/science/article/pii/S0960148112007653.

[12] A. Kusiak, H. Zheng, Z. Song, On-line monitoring of power curves, Renew. Energy 34 (6) (2009) 1487–1493.

[13] R. Bi, C. Zhou, D.M. Hepburn, Detection and classification of faults in pitch-regulated wind turbine generators using normal behaviour models based on performance curves, Renew. Energy 105 (2016) 674–688, http://dx.doi.org/10.1016/j.renene.2016.12.075. http://linkinghub.elsevier.com/retrieve/pii/S0960148116311351.

[14] V. Thapar, G. Agnihotri, V. Sethi, Critical analysis of methods for mathematical modelling of wind turbines, Renew. Energy 36 (11) (2011) 3166–3177.

[15] T. Ouyang, A. Kusiak, Y. He, Modeling wind-turbine power curve: a data partitioning and mining approach, Renew. Energy 102 (2017) 1–8, http://dx.doi.org/10.1016/j.renene.2016.10.032. http://www.sciencedirect.com/science/article/pii/S0960148116308989.

[16] S. Gill, B. Stephen, S. Galloway, Wind turbine condition assessment through power curve copula modeling, IEEE Trans. Sust. Energy 3 (1) (2012) 94–101, http://dx.doi.org/10.1109/TSTE.2011.2167164.

[17] A. Kusiak, A. Verma, Monitoring wind farms with performance curves, IEEE Trans. Sust. Energy 4 (1) (2013) 192–199.

[18] D.C. Montgomery, Introduction to Statistical Quality Control, fourth ed., John Wiley & Sons, 2001.

[19] E. Papatheou, N. Dervilis, A.E. Maguire, I. Antoniadou, K. Worden, A performance monitoring approach for the novel Lillgrund offshore wind farm, IEEE Trans. Ind. Electron. 62 (10) (2015) 6636–6644.

[20] S.W. Doebling, C.R. Farrar, M.B. Prime, D. Shevitz, Damage Identification and Health Monitoring of Structural and Mechanical Systems from Changes in Their Vibration Characteristics: a Literature Review, Tech. rep., Los Alamos National Laboratory LA-13070-MS, 1996.

[21] H. Sohn, C.R. Farrar, F.M. Hemez, D.D. Shunk, D.W. Stinemates, B.R. Nadler, J.J. Czarnecki, A Review of Structural Health Monitoring Literature: 1996-2001, Tech. rep., Los Alamos National Laboratory LA-13976-MS, 2004.

[22] C.R. Farrar, K. Worden, Structural Health Monitoring: a Machine Learning Perspective, John Wiley & Sons, 2013.

[23] K. Worden, G. Manson, N.R.J. Fieller, Damage detection using outlier analysis, J. Sound Vib. 229 (3) (2000) 647–667.

[24] D.A. Clifton, L. Clifton, S. Hugueny, D. Wong, L. Tarassenko, An extreme function theory for novelty detection, IEEE J. Sel. Top. Signal Process. 7 (1) (2013) 28–37.

[25] C.E. Rasmussen, C. Williams, Gaussian Processes for Machine Learning. 2006, vol. 38, The MIT Press, Cambridge, MA, USA, 2006, pp. 715–719.

[26] C. M. Bishop, et al., Neural Networks for Pattern Recognition.

[27] I.T. Nabney, NETLAB: Algorithms for Pattern Recognition, Springer, 2004.

[28] R. Fisher, L. Tippett, Limiting forms of the frequency distributions of the largest or smallest members of a sample, Proc. Camb. Philos. Soc. 24 (1928) 180–190.

[29] E. Castillo, Extreme Value Theory in Engineering, Academic Press, Inc., 1988.

[30] R. Storn, R. Price, Differential evolution a simple and efficient heuristic for global optimisation over continuous spaces, J. Glob. Optim. 11 (1997) 341–359.

[31] K. Worden, G. Manson, H. Sohn, C. Farrar, Extreme value statistics from differential evolution for damage detection, in: Proceedings of the 23rd International Modal Analysis Conference, 2005.

[32] H.W. Park, H. Sohn, Parameter estimation of the generalized extreme value distribution for structural health monitoring, Probabilistic Eng. Mech. 21 (4) (2006) 366–376.

[33] K. Worden, G.R. Tomlinson, Nonlinearity in Structural Dynamics: Detection, Identification and Modelling, IOP Publishing Ltd, 2001.

[34] C.M. Bishop, Neural Networks for Pattern Rcognition, Oxford University Press, 1995.

[35] H. Bourlard, Y. Kamp, Auto-association by multilayer perceptrons and singular value decomposition, Biol. Cybern. 59 (4) (1988) 291–294.

[36] M. Scholz, R. Vigário, Nonlinear PCA: a new hierarchical approach, in: Proc. ESANN, 2002, pp. 439–444.

[37] N. Japkowicz, S.J. Hanson, M.A. Gluck, Nonlinear autoassociation is not equivalent to pca, Neural Comput. 12 (3) (2000) 531–545.

[38] M.A. Kramer, Nonlinear principal component analysis using autoassociative neural networks, AIChE J. 37 (2) (1991) 233–243.

[39] K. Worden, Structural fault detection using a novelty measure, J. Sound Vib. 201 (1) (1997) 85–101.

[40] L. Tarassenko, A. Nairac, N. Townsend, I. Buxton, P. Cowley, Novelty detection for the identification of abnormalities, Int. J. Syst. Sci. 31 (11) (2000) 1427–1439.

[41] D.A. Clifton, S. Hugueny, L. Tarassenko, Novelty detection with multivariate extreme value statistics, J. Signal Process. Syst. 65 (3) (2011) 371–389.