Control Engineering Practice 49 (2016) 204–211

Contents lists available at ScienceDirect

Control Engineering Practice

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

Wind turbine performance analysis based on multivariate higher order moments and Bayesian classifiers

Jürgen Herp [a,*], Niels L. Pedersen [b], Esmaeil S. Nadimi [a]

[a] πSeG The Mærsk Mc-Kinney Møller Institute, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark
[b] Siemens Wind Power A/S, Borupvej 16, DK-7330 Brande, Denmark

ARTICLE INFO

Article history:
Received 29 June 2015
Received in revised form
16 December 2015
Accepted 17 December 2015
Available online 29 December 2015

Keywords:
Wind farm
Multivariate analysis
Bayesian classification
Condition monitoring
k-means clustering

ABSTRACT

A data-driven model based on Bayesian classifiers and multivariate analysis of the power curve (wind speed vs. power) for monitoring wind farms' performance is presented. A new outlier detection criterion and various control bounds on the skewness and kurtosis of the data for cluster separation and classification of turbines' faulty and normal state of operation are introduced. Further continuous monitoring is addressed with Hotelling's T^2 and Bayesian network approaches, and it is proven that under certain conditions, the outcomes of these two methods are equivalent. The Bayesian approach, however addresses the likelihood of classification, making supervised controls more flexible.

1. Introduction

Wind energy is a fast growing renewable energy source, however to make wind power competitive with other sources of energy, availability, reliability and the lifetime of turbines will all need to be improved. As the wind energy sector grows, business economics, will demand increasingly careful management of *Operation and Maintenance* (O&M) costs (Petersen, Madsen, & Bilberg, 2013; Walford, 2006; Wilkinson, Spinato, & Knowles, 2006; Witczak, 2013).

Performance monitoring, especially power curve reconstruction, is assessed in a series of papers (Anahua, Barth, & Peinke, 2008; Gill, Stephen, & Galloway, 2012; Karki & Billinton, 2004; Kusiak, Zheng, & Song, 2009; Rinn, Heißelmann, Wächter, & Peinke, 2013; Schlechtingen, Santos, & Achiche, 2013; Wang, Infield, Stephen, & Galloway, 2013). In Lydia, Kumar, Selvakumar, and Kumar (2014) and Kusiak, Zheng, and Song (2009), various ways to model power curves are summarized. Common for these different approaches are that they deliver good performance under normal operation of a wind farm. However, under abnormal performance, different approaches are applied to asses the new situation. Common prediction, operation, and condition monitoring approaches are summarized in Márquez, Tobias, Pérez, and Papaelias (2012) and Kusiak, Zhang, and Verma (2013). Condition and performance monitoring approaches have shown potential in early fault detection (Arabgol, Ko, & Esmaeili, 2014; Bennouna, Heraud, & Leonowicz, 2012; Fischer, Besnard, & Bertling, 2012; Herp & Nadimi, 2015; Odgaard & Johnson, 2013; Tavner, Xiang, & Spinato, 2007; Yan, Li, & Gao, 2014).

The aim of this paper is to exploit the idea of statistical analysis of wind turbine data instead of physical model based approaches. The main contribution will focus on how a statistical framework can be used to include probabilistic measures in wind turbine controls rather then hard thresholds. It assesses the monitoring of wind turbine performance in a wind farm, based on the power curve. Unlike most performance monitoring the point to point performance of the power curve is not considered. Instead, a multivariate statistical approach to monitor wind turbine performance and detect outliers based on the higher order moments of the multivariate distributions is considered. The approach presented is twofold: (i) due to the stochastic nature of the wind and inherent variability of wind turbines, *Supervisory Control and Data Acquisition* (SCADA) recordings are first processed to extract turbine state information by compressing large data samples to values for the multivariate skewness and kurtosis. The dependence of the overall performance conditioned on an artificial reconstructed reference curve using *k*-means clustering (Lloyd, 2006) and Monte Carlo sampling is considered. This results in identifying and constraining the null hypotheses $\mathcal{H}_0^{(l)}$, $l=1,2$, and selecting a composite hypothesis $\bar{\mathcal{H}}_0$ associated with wind turbine events indicating abnormal behavior. $\mathcal{H}_0^{(1)}$ and $\mathcal{H}_0^{(2)}$ as turbine in a

* Corresponding author.
E-mail address: herp@mmmi.sdu.dk (J. Herp).

http://dx.doi.org/10.1016/j.conengprac.2015.12.018
0967-0661/

normal operating state in which the latter have longer periods of down regulation are defined. Going beyond this simple picture and include fault identification and isolation by benchmark models (Dey, Pisu, & Ayalew, 2015; Odgaard & Stoustrup, 2015; Odgaard, Stoustrup, & Kinnaert, 2013) will not be in the scope of this paper and will be considered in future work. Secondly (ii), the methodology is applied on a wind farm for continuous monitoring and evaluate its performance by introducing probabilistic classifiers.

The paper is organized as follows: Section 2 describes the offshore wind farm data. Section 3 introduces the concept of performance monitoring of wind turbines with respect to the power curve, and describes the prior analysis and filtering for the upcoming multivariate analysis. Sections 4 and 5 describe in detail the multivariate analysis and its outcome. Based on the results of Section 5, a case study is performed by continuously monitoring a wind farm in Section 6. Finally the paper is closed by discussing the outcomes of the investigations in Section 7.

2. Offshore wind farm data

SCADA data are provided from an offshore wind farm comprised of 88 wind turbines (mega Watt scale) that are commissioned in Western Europe in IEC Class IIB climate. The data considered are wind speed v and active power P for each turbine. In addition the turbine log data are available. Assuming all prior events of a wind turbine have been recorded, the turbines can be labelled according to their state of operation. For N samples, the wind turbine data is denoted by

$$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N) \in \mathbb{R}^{2\times N}, \quad N > 2 \tag{1}$$

with x_{1i} containing the wind speed and x_{2i} the associated power measurements, for $i = 1, \ldots, N$. Concerning measurements for January and February 2013 the total number of samples for each turbine is $N=8352$. Each sample is an average over a 10 min interval. Although wind turbine performance depends on the incident wind the assumption is implied that the energy conversion of a wind turbine is independent of the wind direction. Including wind direction in the analysis has shown no impact on the results. Thus, for the sake of simplicity, no conditioning on wind direction is needed to be taken into account. However, in the framework of multivariate statistics additional variable correlations can easily be included by extending $\mathbf{X}$ to more than two sources of measurements.

For the first part of the study, data from January 2013 is chosen, as it resembles best the annual wind speed distribution. The wind farm performance curves for this month are as shown in Fig. 1. Comprising the data from all 88 turbines in the wind farm, Fig. 1 shows the power curve with additional outliers. These make the power curve unsuited to extract moments to represent a clean power curve for future reference when monitoring wind turbines. These outliers are needed to be removed, for which the next section is dedicated.

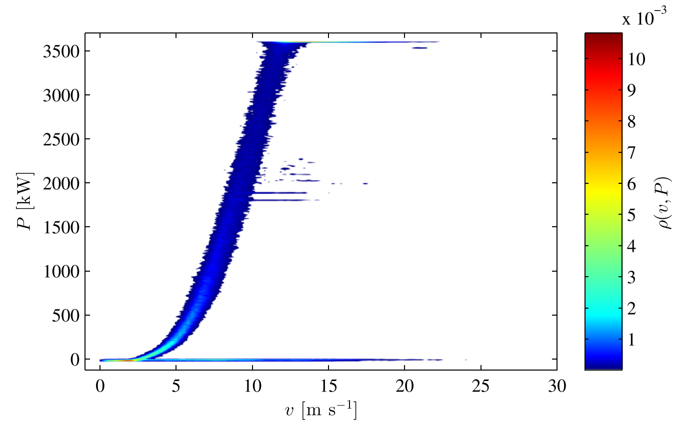

Fig. 1. Raw performance curves: the density plots for the power curve from all 88 turbines. Data indicated with a point (·) exceeds the 99% significance level of the density distribution $\rho(v,P)$.

3. Wind turbine power curves

Performance curves show the relations between multiple variables among each other, take for instance the power curve which relates wind speed to the active power of a wind turbine. Abnormal behavior will have an impact on this relation and influence the capability of a wind turbine to operate optimal. In case of the power curve one can in theory continuously map the wind speed to the power available in the ambient wind stream by means of simple dynamic equations (Burton, Jenkins, Sharpe, & Bossanyi, 2011). However, sensor error, malfunction, respond time etc. impact the relation so that the performance curve gets additional outliers. Other curves for which this holds are the drive train revolution curve, mapping between rotor speed and active power, and the pitch curve, indicating the relation between pitch angle and active power.

The outliers in the power curve will be identified with a k-means clustering algorithm. By applying a distance measure for each data point to a cluster center, data are assigned to a specific cluster. In the case when the distance between a data and its cluster exceeds a threshold the data is removed.

3.1. k-means clustering

A combination of k-means clustering and outlier detection based on the Mahalanobis distance (Mahalanobis, 1936) to obtain representative performance curves is applied. The k-means clustering is a method of vector quantization that is popular in data mining (Bishop, 1995; Zaki & Wagner Meira, 2014). Given the provided data $\mathbf{x}_1, \ldots, \mathbf{x}_N$, k-means clustering aims to partition the N observations into $k \le N$ sets $\mathcal{S} \in \{\mathcal{S}_1, \ldots, \mathcal{S}_k\}$ so as to minimize the total cluster Euclidean distances between each observation:

$$\underset{\mathcal{S}}{\arg\min} \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in \mathcal{S}_j} \|\mathbf{x}_i - C_j\|^2, \quad i = 1, \ldots, N \tag{2}$$

where C_j is the centroid coordinate of the jth cluster. One refer to the function argument in (2), with proper normalization, as the cost function

$$\Delta(\mathbf{X}|k) \equiv \frac{1}{N} \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in \mathcal{S}_j} \|\mathbf{x}_i - C_j\|^2. \tag{3}$$

The Euclidean distance does not contain information on the variance or the separation of the clusters. In order to prepare for over fitting on the reference curves, one consider other measures for the optimum number of clusters as well. The Calinski–Harabasz criterion (Caliński & Harabasz, 1974) describes the overall variation of the clusters with respect to the sample mean $\boldsymbol{\mu}$:

$$\Xi(\mathbf{X}|k) \equiv \frac{N-k}{k-1} \frac{\sum_{j=1}^{k} n_j \|C_j - \boldsymbol{\mu}\|^2}{N\Delta(\mathbf{X}|k)}, \tag{4}$$

where n_j is the number of samples in the jth cluster.

In order to confine the numbers of clusters further, a new measure for how well each sample lies within a cluster is