Wind Turbine Performance Degradation Monitoring Using DPGMM and Mahalanobis Distance

Peng Guo[1], Yu Gan[1], David Infield[2]

(1. the School of Control and Computer Engineering, North China Electric Power University, Beijing,102206, China.

2. the Institute of Energy and Environment within the Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, G1 1XW, UK)

Abstract

Real-time monitoring of wind turbine performance degradation can improve the economics and safety of wind farms. Normal operational data can accurately reflect the generation performance of a wind turbine and in the wind-speed coordinate system these normal data constitute the “main power band”. This paper invokes a Dirichlet Process Gaussian Mixture Model (DPGMM) to cluster operational data in each horizontal power bin, and the number of Gaussian components can be determined automatically. The confidence ellipses of Gaussian components can be used to identify the contour of the main power band which is then used as baseline performance model. In the monitoring phase, Mahalanobis distance is used to judge whether new monitoring data lies outside the contour of main power band and thus should be labeled as degraded operational data. When the proportion of such data exceeds a set value in a sliding window, a wind turbine performance degradation alarm is triggered. Degradation degree and rate can quantitatively measure the severity of performance degradation. For an industrial performance degradation case caused by gearbox oil over temperature, the method proposed timely gives alarm only 12 points (2 hours) later than the first

This is a peer reviewed, accepted author manuscript of the following research article: Guo, P., Gan, Y., & Infield, D. (2022). Wind turbine performance degradation monitoring using DPGMM and Mahalanobis distance. Renewable Energy, 200, 1-9. https://doi.org/10.1016/j.renene.2022.09.115

degraded operational data appears and is proved to be effective.

Keywords: wind turbine; performance degradation; Dirichlet Process Gaussian Mixture Model (DPGMM); Mahalanobis distance.

1. Introduction

Some component faults and abnormalities may not result in the shutdown of wind turbines or trigger alarms, but can significantly reduce the output power and so degrade wind turbine performance. Real-time analysis of SCADA data to detect performance degradation in a timely manner can improve the safety and economic benefits of wind turbines.

The power curve is an important measure of the generation performance of a wind turbine [1-3]. The method of bins is used to create the power curve. Ten-minute averaged SCADA data are allocated to different wind speed bins according to their wind speed, each bin having a wind speed interval of 0.5m/s. Average wind speed and average power are calculated for each wind speed bin and these averaged pairs comprise the power curve by joining the pairs sequentially, from lowest to highest wind speed. But the averages absorb too much data variation and comparing power curves alone may not provide sufficient quantitative information about performance differentiation or degradation. [4,5] carried out comprehensive review and comparison of different wind turbine power curve modeling methods. [6,7] used machine learning method such as Gaussian Process (GP), Random Forest (RF), Support Vector Regression (SVR) and k-Nearest Neighbors (KNN) to construct multivariable power curve model and detected performance degradation through analysis of power predicting residuals. In [8], the authors applied a two-phase method for assessing wind turbine performance. In the first phase, wind power is predicted by an ensemble of extreme learning machines. In the second phase, the predicted power and wind speed are used to construct a Copula model and parameters of the Copula models served as metrics for assessing the performance

of wind turbines. [9] constructed linear and Weibull based power curve models and used a control chart as residual analysis methods to monitor the wind turbine generation performance. [10] proposed a daily performance monitoring method for wind turbines. After data cleaning with k-mean method, five-parameters Logistic regression function was used to model the power curve with cleaned data. Improved fuzzy comprehensive evaluation method was established to monitoring future wind turbine condition. But methods in [6-10] for wind turbine performance monitoring are not really intuitive, and how to quantitatively measure the degree of performance degradation is still not properly resolved. [11] evaluated the wind turbine performance by performing principal component analysis on the quasi-linear region of power curve and used the standard deviation of the secondary principal component as health value for performance degradation assessment. But the power curve between cut-in and rated wind speeds is related to the cubic of wind speed and far from linear shape which may lower the accuracy of degradation monitoring result. [12] predicted the active power via ensembling of multivariate polynomial regression models that exploit a higher number of input (include environmental variables and operational variables) for performance analysis, but did not give detailed method for degradation judgement. In [13], the authors put forward two methods for creating power threshold curves that was used to monitor performance degradation such as blade ice accretion. The first method relies on a percentage deviation from the manufacturer's power curve. The second method obtained threshold curve based on the observed variance of operational data. When monitoring data consecutively occurred outside the power threshold, degradation such as blade icing may happen. But the power threshold curve was greatly influenced by the data variance which reduced the accuracy of monitoring results.

In this paper, an intuitive and effective performance degradation monitoring method is proposed. Contour of main power band constituted with normal operational data is intuitively extracted and used as performance model. Deviation of monitoring data from performance model is quantitatively measured by Mahalanobis distance.

Degradation degree and rate can be accurately calculated to give degradation alarms. The effectiveness of the method is demonstrated with an industrial study case. The content of the paper is organized as follows. Section 2 outlines the performance degradation monitoring principle. In Section 3, the wind turbine performance model is constructed using a DPGMM. Subsequently in Section 4, the Mahalanobis distance and sliding window are invoked to generate performance degradation alarms.

2. Data from Test Wind Turbine and Performance Degradation Monitoring Principle

The test wind turbine is rated at 1.5MW and named A03. The SCADA data is 10-min averaged and covers 63 variables including wind speed, active power, pitch angle, gearbox temperature and ambient temperature. There are totally 4320 records from 1/5/2019 to 30/5/2019 as shown in Fig.1. The wind speed and power in each record comprise a data point in the wind speed-power (V-P) coordinate system. Data points from normal operation during 1/5 to 28/5 are located densely to form the so called "main power band". The characteristics of the main power band such as the envelope contour and the distribution of the data points are a real reflection of the wind turbine generation performance and can be used as a baseline against which to determine any degradation. During the period 29/5 to 30/5, towards the end of the monitoring, the test wind turbine exhibits obvious degradation. Many data points with power values well below the normal points at same wind speed lie outside the right edge of the main power band. More isolated individual points outside the main power band are mainly the result of sensor failure, or reflect a transient process of wind turbine startup or shut down. However, data points like those from 29/5 to 30/5 appeared consecutively and with an obviously horizontal distribution characteristic, it suggests that the wind turbine experienced performance degradation caused by some abnormality or a failure of components.

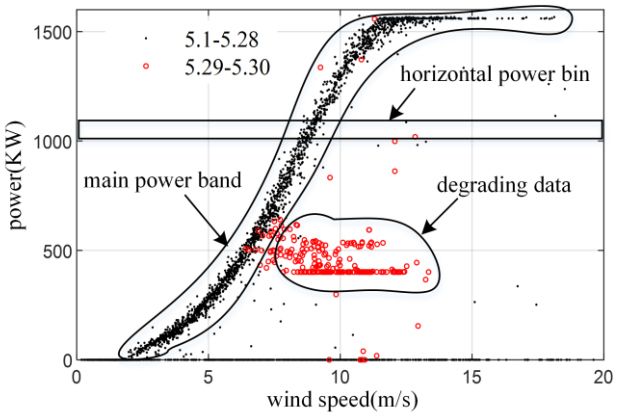


Fig.1 raw SCADA data from test wind turbine

This paper extracts the contour of main power band in Fig.1 as the wind turbine baseline performance model using a horizontal stratified Dirichlet Process Gaussian Mixture Model (DPGMM). After the performance model is constructed, the Mahalanobis distance is used to provide a timely and intuitive measure of monitoring data deviation from the performance model. When degraded operational data occur consecutively, performance degradation alarms are triggered.

With the passage of time, the performance model should be updated with new SCADA data to ensure model accuracy and reflects the fact that wind turbine performance does slowly change as result of ambient factors such as air density. A cyclic performance degradation monitoring strategy composed of a model construction phase and monitoring phase is implemented. Details of this strategy can be found in [14].

3. Wind Turbine Generation Performance Model Construction

3.1 Procedures for the construction of wind turbine performance model

For the data of 1/5 to 28/5, besides normal data, there are some abnormal data lie outside the main power band. In order to extract the contour of main power band as performance model, the abnormal data should be firstly identified and removed. In this paper, DPGMM is used to construct the performance model with following two steps:

(1) Creating horizontal power bins. As shown in Fig. 1, the distribution of data points in V-P coordinate system is complex, and it is difficult to extract the contour of main

power band as a whole. Because the data points above the rated wind speed and those with artificial power curtailment show horizontal distributions, creating multiple power bins with a certain interval (i.e., in the horizontal power direction) and analyzing data points in each power bin successively can provide useful information. In order to facilitate data processing and ensure sufficient data points in each power bin, the power interval is set as 50KW in this paper.

（2）DPGMM is used to cluster the data point in each power bin for identifying and removing the abnormal data. Confidence ellipses of Gaussian components of DPGMM for clustering normal data in each power bin form the contour of main power band which is used as performance model.

3.2 Data cluster for power bin with Dirichlet Process Gaussian Mixture Model

For each power bin below rated power, DPGMM [15-18] is used to identify abnormal data and extract the contour of main power band. Compared to a traditional Gaussian Mixture Model which needs the clustering number fixed beforehand, a DPGMM which is an infinite mixture model can automatically determine the clustering number according to distribution characteristics of the data and thereby get an improved clustering result. The Dirichlet Process (DP) provides prior distribution parameters for clusters of DPGMM and translates a finite clustering problem into an infinite one. For DPGMM, the cluster number can increase adaptively to the complexity of the data. Clusters number and parameters for DPGMM can be obtained through iteration with Chinese Restaurant Process (CRP) or stick-breaking process methods.

A Dirichlet Process (DP) is a stochastic process that defines a probability distribution on an infinite dimensional space. For an arbitrary segmentation of sample space as $A=\bigcup_{i=1}^{n} A_i$, if distribution G has characteristics as (1),

$$\begin{aligned}&\left(G(A_1),G(A_2),\cdots,G(A_n)\right)\sim\\&\mathrm{Dir}\left(\alpha G_0(A_1),\alpha G_0(A_2),\cdots,\alpha G_0(A_n)\right)\end{aligned}\tag{1}$$

then G follows a Dirichlet Process (DP) as:

$$G\sim DP(\alpha,G_0)\tag{2}$$

where, α is a concentration parameter that controls the probability of creation of new

cluster during iteration (the greater α is, the higher the probability of creating new cluster), G_0 is base distribution that defines the initial shape of the clusters, and Dir is a Dirichlet distribution.

In a Dirichlet Process Mixture Model (DPMM), DP serves as prior distribution for the parameters of clusters.

Assume that there are N data points in a power bin as:

$$\boldsymbol{D} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_N] \tag{3}$$

In DPMM, the probability $p(\boldsymbol{x}_i)$ for a data point $\boldsymbol{x}_i$ in a power bin follows:

$$\begin{cases} p(\boldsymbol{x}_i) = \sum_{k=1}^{\infty} \pi_k F(\boldsymbol{\theta}_k); \\ \boldsymbol{\theta}_k \mid G \sim G; \\ G \sim DP(\alpha, G_0) \end{cases} \tag{4}$$

where, $F(\boldsymbol{\theta}_k)$, $\boldsymbol{\theta}_k$ and π_k are respectively the distribution function, distribution parameters and weight for cluster k. And the infinite number of clusters learned through DP iteration can converge to be finite as K. Because in this paper we need to cluster two-dimensional data points with wind speed and power, two-dimensional Gaussian distribution is chosen for $F(\boldsymbol{\theta}_k)$. A DPMM with Gaussian distribution as cluster is called Dirichlet Process Gaussian Mixture Model (DPGMM) and each Gaussian distribution for the mixture model is called a Gaussian component. $\boldsymbol{\theta}_k = [\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k]$ is the parameter for the Gaussian component with mean value $\boldsymbol{\mu}_k$ and variance $\boldsymbol{\Sigma}_k$. The distribution of $\boldsymbol{\theta}_k$ follows G which has base distribution G_0 and concentration parameter α.

In order to determine the number (K) of Gaussian components of DPGMM in each power bin, cluster weights π_k and the parameters $\boldsymbol{\theta}_k$ for each component, following posterior probability for DPGMM will be calculated.

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_K, \pi_1, \pi_2, \cdots, \pi_K \mid \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_N) \tag{5}$$

In this paper, a Gibbs sampler [19-20] is used to calculate the posterior probability of (5).

Data clustering with DPGMM for each power bin is shown in Fig.2.

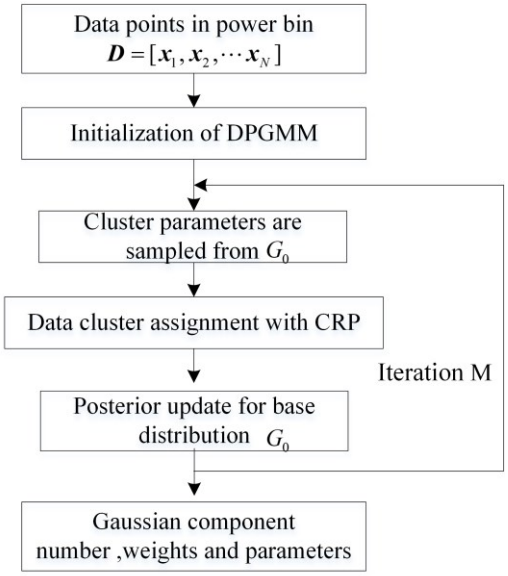


Fig.2 Procedure for DPGMM clustering in each power bin

As shown in Fig.2, the procedure for data clustering with DPGMM in each power bin are as following:

(1) Initialization of DPGMM for a power bin. Iteration number $M=200$, concentration parameter $\alpha=20$. Because the DPGMM can adaptively increase the cluster number according to data complexity, the initial cluster number can be small as 2. The base distribution G_0 is selected as a Normal Inverse Wishart (NIW) that has conjugate relationship with Gaussian distribution as (6):

$$G_0 = NIW(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \gamma, \nu) \tag{6}$$

$\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \gamma, \nu$ are parameters for NIW. The value for the G_0 can be initialized as:

$$\boldsymbol{\mu}_0 = \sum_{i=1}^{N} \boldsymbol{x}_i / N, \boldsymbol{\Sigma}_0 = \boldsymbol{D} \cdot \boldsymbol{D}^{\mathrm{T}}, \gamma = \nu = N \tag{7}$$

(2) The parameters for cluster $\boldsymbol{\theta}_k$ are sampled from the base distribution G_0 and the prior Dirichlet distributions are determined.

(3) Cluster assignment for each data point $\boldsymbol{x}_i$ in the power bin is implemented with Chinese Restaurant Process (CRP) in Fig.3. CRP is a classical implementation approach for Dirichlet Process. CRP describes DP as a customer (each data point) sitting (cluster assignment) at a table (cluster) in a restaurant as shown in Fig.3. The restaurant is infinitely large, and there are enough tables to place infinitely. Customers enter the restaurant one by one, and each customer observes the current table situations before

deciding which table to sit on. The more seated the table is, the more likely it is to be a popular table. A new table is placed at an arbitrary position each time the new customer enters and is kept or discarded according to the customer's decision. The probabilities that the new customer (data point) will sit at the existing table (existing cluster) k and the new table (new cluster) are shown as (8).

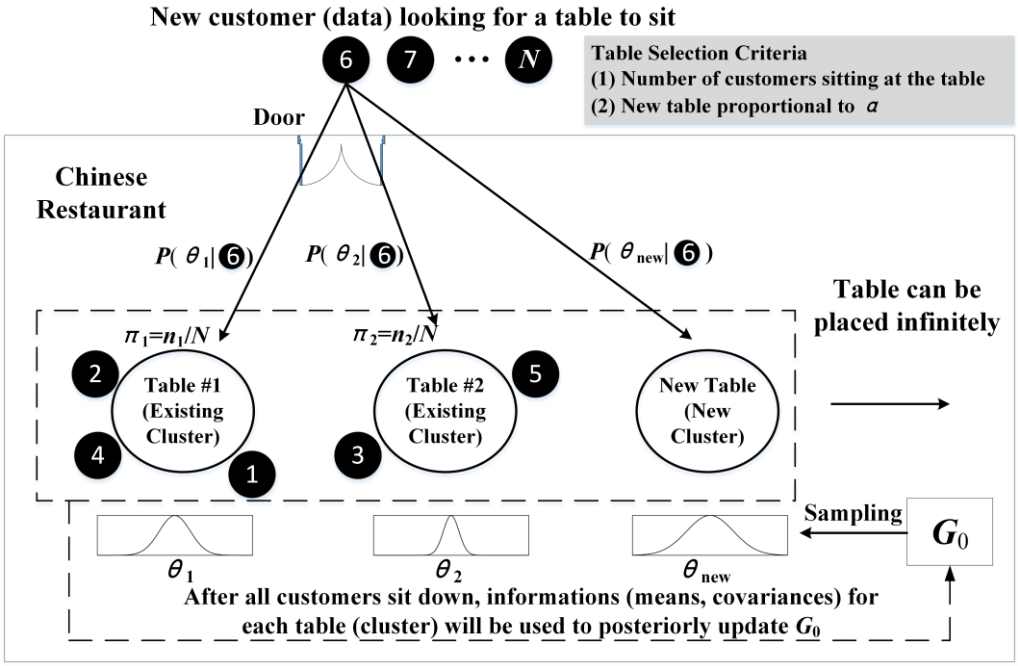


Fig.3 Chinese Restaurant Process (CRP)

$$\begin{cases} P(\boldsymbol{\theta}_k \mid \boldsymbol{x}_i) = \dfrac{n_k}{N-1+\alpha} F(\boldsymbol{x}_i, \boldsymbol{\theta}_k) \\ P(\boldsymbol{\theta}_{\text{new}} \mid \boldsymbol{x}_i) = \dfrac{\alpha}{N-1+\alpha} F(\boldsymbol{x}_i, \boldsymbol{\theta}_{\text{new}}) \end{cases} \tag{8}$$

where, N is the customer number (data point number), $P(\boldsymbol{\theta}_k \mid \boldsymbol{x}_i)$ is the probability of new customer (data point) $\boldsymbol{x}_i$ being assigned to an already existing table (cluster) k with parameter $\boldsymbol{\theta}_k$; $F(\boldsymbol{x}_i, \boldsymbol{\theta}_k)$ is the probability of $\boldsymbol{x}_i$ under the Gaussian distribution with parameter $\boldsymbol{\theta}_k$; n_k is the customer number already sitting at table (cluster) k. $P(\boldsymbol{\theta}_{\text{new}} \mid \boldsymbol{x}_i)$ is the probability of new customer $\boldsymbol{x}_i$ being assigned to a new table (cluster) with parameter $\boldsymbol{\theta}_{\text{new}}$; $F(\boldsymbol{x}_i, \boldsymbol{\theta}_{\text{new}})$ is the probability of $\boldsymbol{x}_i$ under new Gaussian distribution with parameter $\boldsymbol{\theta}_{\text{new}}$.

From (8), the probability of a new customer (data point) sitting at an existing table (cluster) is proportional to the number of customers already at that table. While the

probability of a new table tends to be proportional to the concentration parameter. For this reason, the DP is also metaphorically referred to as “richer-get-richer” approach.

Based on the probabilities of being assigned to different tables (clusters), a multinomial distribution with number of trials fixed as 1 is used to finally determine which table (cluster) the customer (data point) $\boldsymbol{x}_i$ belongs to as follow:

$$Cluster_i \sim \text{Multi}(1, P(\boldsymbol{\theta}_k \mid \boldsymbol{x}_i)) \tag{9}$$

where, $\text{Multi}()$ is a multinomial distribution; $Cluster_i$ is the cluster $\boldsymbol{x}_i$ belongs to; $P(\boldsymbol{\theta}_k \mid \boldsymbol{x}_i)$ is the probability of $\boldsymbol{x}_i$ being assigned to cluster k with parameter $\boldsymbol{\theta}_k$.

After cluster assignment for each data point in the power bin, clusters with zero data point will be discard. Assume that the number of clusters which has data point is K' and these clusters respectively have $n_1, n_2, \cdots, n_k, \cdots, n_{K'}$ data point. For cluster k, it’s parameter $\boldsymbol{\theta}_k$ is sampled from base distribution G_0, and its weight in (4) can be calculated as:

$$\pi_k = \frac{n_k}{N} \tag{10}$$

(4) After cluster assignment, the base distribution G_0 is updated posteriorly and the NIW parameters in (6) can be recalculated with data in each cluster according to [15].

With iteration of steps (2) to (4), the final cluster number K, weight π_k and parameters $\boldsymbol{\theta}_k = [\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k]$ for Gaussian components can be obtained when the iteration converges. After the DPGMM is constructed, for a data point in a power bin, its probability is the weighted sum of its probabilities belonging to K Gaussian components as:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k F_k\left(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \tag{11}$$

where, π_k is the weight for Gaussian component k which reflects the importance of this component during clustering, and $\sum_{k=1}^{K} \pi_k = 1$. The program for DPGMM data clustering was implemented using Python 3.6 with libraries of Numpy and Scipy.

3.3 DPGMM clustering analysis for power bins

For the test wind turbine, two representative power bins of 600-650KW and 950-

1000KW are selected for investigation. In power bin 600-650KW, there are no abnormal data, but there are isolated abnormal data in power bin 950-1000KW.

Data clustering with DPGMM in power bin of 600-650KW is shown in Fig.4. The mean values and weights for Gaussian components are listed in Table.1.

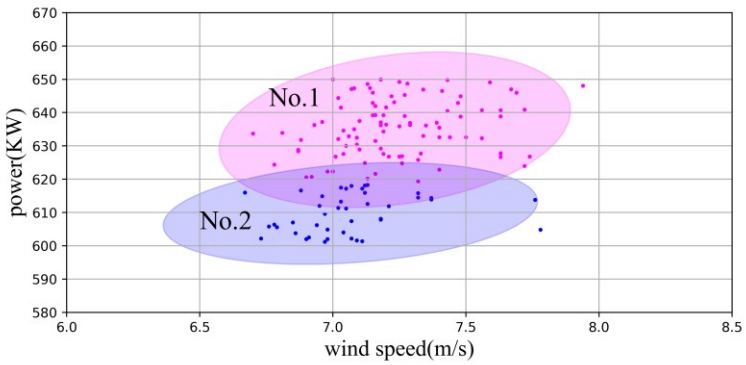


Fig.4 DPGMM clustering for 600-650KW power bin

Table.1 DPGMM parameters for 600-650KW power bin

component	weight	mean value
No.1	0.699	[**7.23m/s**, 634.8KW]
No.2	0.301	[**7.07m/s**, 609.7KW]

In Fig.4, DPGMM accurately clusters data points in 600-650KW power bin with two Gaussian components. The confidence ellipse of each Gaussian component is plotted and the confidence level is the commonly used value 95%. As a result, two data points far away from the ellipse centers are not included in the ellipses. The center for confidence ellipse is mean value $\boldsymbol{\mu}_k$ for the Gaussian distribution which is composed of mean wind speed and mean power. In this power bin, weights for these two components are large reflecting that they all cluster considerable number of data points. And the mean wind speeds of these two components are quite similar around 7m/s which show that these two confidence ellipses are located quite closely in Fig.4.

Fig.5 is the clustering result for power bin 950-1000KW with DPGMM. And parameters for each Gaussian component are shown in Table.2.

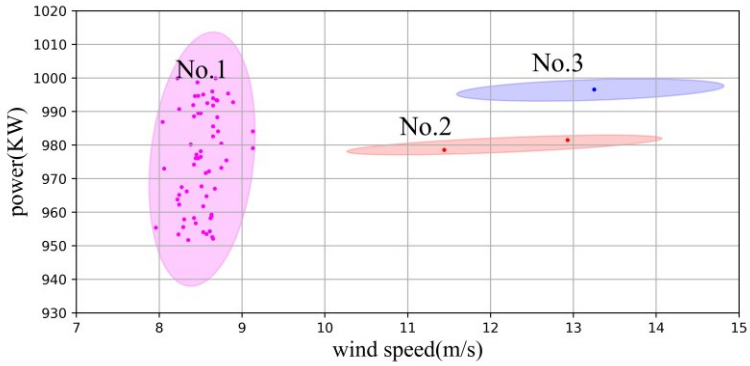


Fig.5 DPGMM clustering for 950-1000KW power bin

Table.2 DPGMM parameters for 950-1000KW power bin

component	weight	Mean value
No.1	0.93	[**8.52m/s**, 975.81KW]
No.2	0.03	[**13.20m/s**, 996.40KW]
No.3	0.04	[**12.17m/s**, 980.03KW]

In 950-1000KW power bin, there are some isolated abnormal data. In Fig.5, DPGMM automatically uses 3 Gaussian components for clustering. Gaussian component 1 clusters normal data points and has large weight as 0.93. Components 2 and 3 cluster the isolated abnormal data points with very small weights of 0.03 and 0.04. In addition, the mean wind speeds for component 2 and 3 are respectively 13.20m/s and 12.17m/s which are significantly larger than the mean wind speed of component 1 whose value is only 8.52m/s. Fig.5 shows that in this instance the confidence ellipses of component 2 and 3 which cluster abnormal data are quite distant from the confidence ellipse of component 1.

3.4 Extracting the contour of main power band and performance model construction

As shown in Fig.4 and Fig.5, DPGMM can accurately cluster data points in power bins. In order to distinguish normal Gaussian component clustering normal data points and abnormal Gaussian component clustering abnormal data, following method is adopted.

Comparing with normal data in a power bin such as 950-1000KW in Fig.5, abnormal data has a much higher wind speed while with a similar power that make them lie at the right side and far away from the main power band. With this reason, in

a power bin, the confidence ellipses of Gaussian components that cluster the normal data points are always located to the left side of the collection of confidence ellipses. While components cluster abnormal data are located at the right side of the power bin and far away from the normal ones. Such as in Fig.5 for 950-1000KW power bin, the normal confidence ellipse of Gaussian component 1 with mean wind speed of 8.52m/s is located at the left side of the power bin, while the abnormal confidence ellipses of components 2 and 3 with mean wind speed respectively of 13.2m/s and 12.17m/s are quite far away from normal component 1. In Fig.4 of 600-650KW power bin, two normal confidence ellipses of Gaussian component 1 and 2 are locate quite closely with mean wind speed respectively as 7.07m/s and 7.23m/s.

With above analysis, among all Gaussian components of the DPGMM in a particular power bin, the Gaussian component with the smallest mean wind speed is selected as base normal Gaussian component, such as component 2 in Fig.4 and component 1 in Fig.5, and its mean wind speed is denoted as $GBase$. Define a center range value for the confidence ellipse of Gaussian component as G_{V}. In a power bin, if the mean wind speed of a Gaussian component is smaller than $GBase+G_{\mathrm{v}}$, this component will be located close to the base normal Gaussian component and is also labeled as normal component. Otherwise, the component will be located at a distance from the base normal Gaussian component and is labeled as an abnormal component. G_{V} reflects the range of centers for normal Gaussian components in a power bin and is decided as follows.

The relationship between the power and wind speed of a wind turbine is:

$$P=\frac{1}{2}\pi C_{\mathrm{p}}\rho R^2V^3 \tag{12}$$

where, V is wind speed, C_{p} is the power coefficient, ρ is the air density, and R is the rotor radius.

Therefore, in a power bin with an interval ΔP, the expected variation range of wind speed ΔV is:

$$\Delta V = \frac{\Delta P}{\frac{3}{2}\pi C_{\mathrm{p}} \rho R^2 V^2} \tag{13}$$

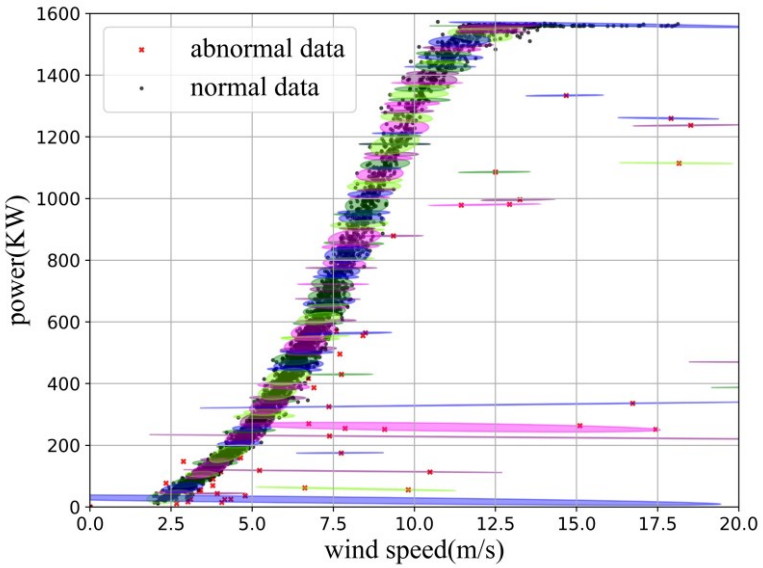
For the test turbine, the wind speeds corresponding to cut-in and rated are 3m/s and 10m/s respectively. The rotor radius is 48 meters, and the air density is 1.225kg/m^3. The average value of the turbine power coefficient below rated wind speed is 0.4, calculated based on the operating data. When the power bin interval is 50KW, the corresponding wind speed variation range for the power bin at cut-in and rated wind speed are $\Delta V_{\mathrm{cutin}} = 1.04\mathrm{m/s}$ and $\Delta V_{\mathrm{rated}} = 0.094\mathrm{m/s}$ respectively. The wider wind speed variation range (i.e., at the cut-in wind speed) $\Delta V_{\mathrm{cutin}} = 1.04\mathrm{m/s}$ is taken as the reference, G_{V} is determined as:

$$G_{\mathrm{v}} = 1.1\mathrm{m/s} \tag{14}$$

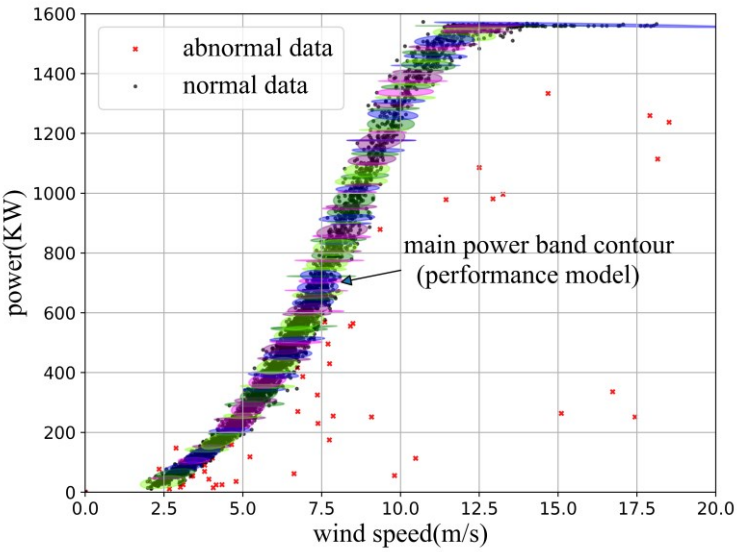
With above method for identifying normal and abnormal Gaussian components, in power bin 600-650KW, component 2 is the base normal Gaussian component, and $GBase = 7.07$ m/s. The mean wind speed of Gaussian component 1 is 7.23m/s that is smaller than $GBase + G_{\mathrm{v}} = 8.17$ m/s, therefore component 1 is also labeled as normal.

In power bin of 950-1000KW, component 1 is the base normal Gaussian component with $GBase = 8.52$ m/s. The mean wind speeds of components 2 and 3 are respectively 13.20m/s and 12.17m/s and both obviously greater than $GBase + G_{\mathrm{v}} = 9.62$ m/s. And components 2 and 3 are labeled as abnormal. Data points clustered by abnormal Gaussian component such as 2 and 3 are also labeled as abnormal data.

The DPGMM clustering and abnormal data labeling method explained above are used for each power bin below rated power and the result is shown in Fig.6.



(a) Abnormal data clustering and identification



(b) The contour of main power band

Fig.6 Data clustering and contour of main power band

In Fig.6(a), for each power bin, DPGMM method correctly clustering data according to dense or sparse data distributions. And abnormal data and corresponding confidence ellipses are correctly identified. In Fig.6(b), normal Gaussian components of the DPGMM for each power bin cluster the normal data points. The contour of main power band which is composed of normal data is determined by the confidence ellipses (in different colors) of the normal Gaussian components in each power bin, such as components 1 and 2 in 600-650KW power bin or component 1 in 950-1000KW power

bin. Abnormal data points clustered by abnormal Gaussian components in Fig.6(b) all lie outside the contour of main power band.

The baseline generation performance model for wind turbine is the contour of main power band which is comprised of the confidence ellipses of normal Gaussian components in each power bin.

4. Wind Turbine Performance Degradation Monitoring

4.1 Degradation assessment of monitoring data using Mahalanobis distance

In the monitoring phase, after the baseline generation performance model has been constructed as in Section 3.4, if a new monitoring data point lies inside the contour of main power band, that is, within the confidence ellipses of normal Gaussian components, the new monitoring data point will be identified as a normal data. Otherwise, if a monitoring data lies right outside the confidence ellipses of the normal Gaussian components, compared with normal data in the contour of main power band of same power bin, the data has an obviously higher wind speed while with a similar power, it will be defined as a degraded operational data.

In order to accurately identify such data in relation to the confidence ellipses of the normal Gaussian components, Mahalanobis distance [21-23] is introduced. Mahalanobis distance can measure the distance between a point $\boldsymbol{x}$ and a distribution P as:

$$d_{\mathrm{m}}(\boldsymbol{x}) = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \tag{15}$$

where, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are respectively mean value and variance matrix for distribution P. In this paper, $\boldsymbol{x}$ is the monitoring data point, and P is the normal Gaussian component in each power bin.

Take power bin 600-650KW and new monitoring data point $\boldsymbol{x}_{\mathrm{obs}} = [8.72\mathrm{m/s}, 642.2\mathrm{KW}]$ as an example. The following steps are needed for assessing whether $\boldsymbol{x}_{\mathrm{obs}}$ is a degraded operational data.

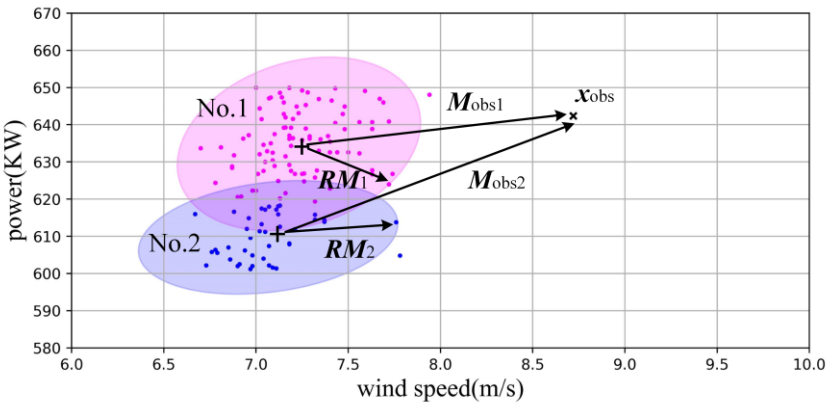


Fig.7 Degradation judgment for monitoring data

(1) Decide which power bin the new monitoring data belonging to and calculate the Mahalanobis radius for the confidence ellipses of the normal Gaussian components, denoted as RM in this power bin. Because the power is 642.2KW, this data belongs to power bin 600-650KW. RM can be calculated as the maximum Mahalanobis distance of all data points in the confidence ellipse to the ellipse center $\boldsymbol{\mu}$. As shown in Fig.7, there are two normal Gaussian components 1 and 2 in 600-650KW power bin, and the RM_1 and RM_2 are the distances of data points on the edge of confidence ellipses to the centers.

(2) Calculate the Mahalanobis distances between the monitoring data and centers of the confidence ellipses as M_{obs}. As shown in Fig.7, distances between monitoring data $\boldsymbol{x}_{\text{obs}}$ and centers of confidence ellipses 1 and 2 are respectively M_{obs1} and M_{obs2}.

(3) If the distances between monitoring data and centers of confidence ellipses are greater than the Mahalanobis radius of confidence ellipse, that is, $M_{\text{obs1}} > RM_1$ and $M_{\text{obs2}} > RM_2$, then the new monitoring data lies outside the contour of the main power band as shown in Fig.7, this indicates degradation.

Table.3 shows the Mahalanobis radius and distance between the monitoring data to the confidence ellipse centers.

Table.3 Mahalanobis radius and distance for monitoring data in 600-650KW power bin

component	No.1	No.2
center	[7.23m/s, 634.8KW]	[7.07m/s, 609.7KW]
RM	2.54（RM_1）	2.55（RM_2）
M_{obs}	5.76（M_{obs1}）	7.27（M_{obs2}）
M_{obs} / RM	2.27	2.85

In Table.3, the ratio M_{obs} / RM between Mahalanobis distance M_{obs} for $\boldsymbol{x}_{\text{obs}}$ and the Mahalanobis radius of confidence ellipse RM can accurately describe how far the monitoring data is away from the main power band. If M_{obs} / RM is greater than 1, it means that the monitoring data lies outside the corresponding confidence ellipse. In this paper, degradation degree is defined as:

$$Degrade = \min\left\{M_{\text{obs}}(i) / RM(i), i = 1, 2, \cdots, n\right\} \tag{16}$$

where, $RM(i)$ and $M_{\text{obs}}(i)$ are respectively the Mahalanobis radius and Mahalanobis distance of monitoring data to confidence ellipse of $i-\text{th}$ normal Gaussian component; n is the number of normal Gaussian components of the power bin where the monitoring data lies. In 600-650KW power bin, the degradation degree of $\boldsymbol{x}_{\text{obs}}$ is:

$$Degrade_{\text{xobs}} = \min\left\{M_{\text{obs}}(1) / RM(1), M_{\text{obs}}(2) / RM(2)\right\} = \min\left\{2.27, 2.85\right\} = 2.27 \tag{17}$$

Monitoring data with a larger degradation degree will be further away from the main power band, reflecting that the degradation is more serious.

In order to ensure that the degraded operational data notably locating at the right side of the main power band, a degradation value is set as: $V_{\text{degrade}} = 1.1$. New monitoring data with a degradation degree $Degrade_{\text{xobs}} > V_{\text{degrade}}$ will be labeled as degraded operational data such as $\boldsymbol{x}_{\text{obs}}$.

The concentration parameter α is the hyperparameter for DPGMM. As mentioned before, a larger α will lead to a higher probability of creating new cluster (Gaussian component) for data points in a power bin. In above studies, α is set as 20. For a comparison, DPGMM with a larger hyperparameter α as 40 is applied to power bin of 600-650KW and the result is shown in Fig.8.

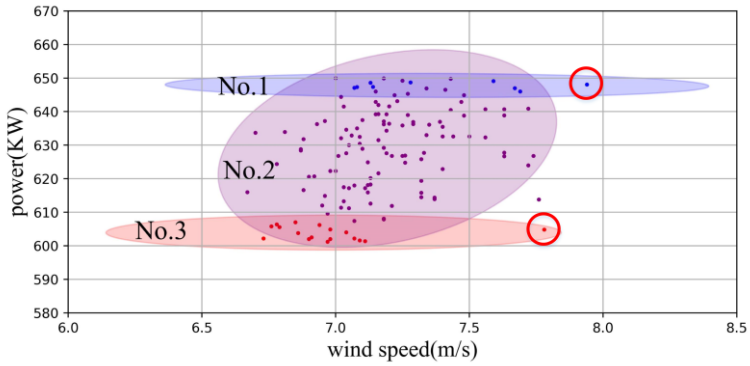


Fig.8 DPGMM with larger hyperparameter clustering for 600-650KW power bin

Comparing Fig.4 and Fig.8, we can see that with larger hyperparameter α, DPGMM tends to create more components of three (No.1 to No.3) in Fig.8 instead of two (No.1 to No.2) components in Fig.4 to cover the isolated data points. This will result two disadvantages for degradation monitoring: (1) Increase of components will add computing burden and cost to degradation monitoring; (2) because more components (such as No.1 and No.3) in Fig.8 tend to cluster isolated data point in the power bin, the contour of the main power band (confidence ellipses of No.1 and No.3) will be erroneously expanded to cover the isolated data points which will decrease the sensitivity for performance degrading monitoring comparing to a relatively narrow contour in Fig.4. With above reasons, smaller hyperparameter α is recommended.

4.2 Creating a robust wind turbine performance degradation alarm

As explained earlier, sensor failure and transient processes of wind turbine startup or shutdown may randomly and intermittently produce data that could suggest degraded operation. In order to detect systematic performance degradation in a timely manner and reduce the number of false alarms, a sliding window [24] monitoring method is adopted to trigger alarms. The window width is N_{win}, that is, there are N_{win} monitoring data in the sliding window. Each sliding window will be updated with N_{update} new data as shown in Fig.9. It should be noted that the monitoring data in a sliding window should have power value greater than zero.

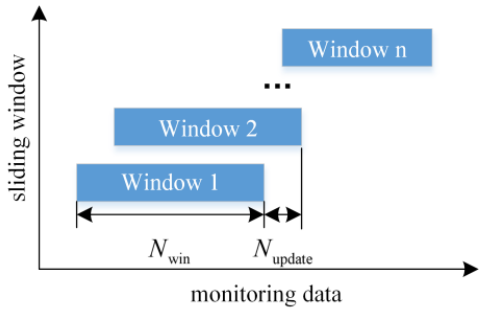


Fig.9 Sliding windows

The method of Section 4.1 is used to judge whether the N_{win} monitoring data in the sliding window are degraded operational data, one by one. Assuming there are $N_{degrade}$ such data in a sliding window, then the degradation rate can be defined as:

$$R_{degrade} = \frac{N_{degrade}}{N_{Win}} \times 100\% \tag{18}$$

Set a wind turbine performance degradation alarm value as V_{alarm}. If the degradation rate of a sliding window exceeds this value, a performance degradation alarm is triggered. The alarm moment is the time stamp of the latest monitoring data in the sliding window.

4.3 Industrial study case of wind turbine performance degradation

As shown in Fig.6(b), the wind turbine baseline performance model is constructed with data from 1/5 to 28/5. Data after 0:00 29/5 are used as monitoring data. Sliding window with large width (data number) will have better average effect which can suppress random noise and reduce false alarm but may lead to low alarm sensitivity. Because the SCADA data are 10-min averages which itself is an average for 600 1-second sampling raw data, the window width and updating data number is defined moderately as 30 (5 hours) and 6 (1 hour) that can reach a satisfied balance between reliability and sensitivity. The degradation alarm value is set as V_{alarm}=40%. Fig.10 shows the degradation rate trend from 0:00 29/5. Fig.11(a) shows the degraded operational data in 5th sliding window. Fig.11(b) shows all such data during 29/5-30/5. Compared with the raw SCADA data in Fig.1, the method of combining DPGMM and Mahalanobis distance accurately identifies degraded operational data which lies outside

the contour of the main power band (performance model).

In Fig.10, the first degraded operational data appears in the 3th window, and degradation rates of next sliding windows gradually increase. At the 5th sliding window, the degradation rate reaches 43.3% and exceeds the alarm value (V_{alarm}=40%), that is, at $N_{\text{win}}+4\times N_{\text{update}}=54$ monitoring data of 9:00 29/5, performance degradation alarm for the test wind turbine is triggered. This alarm is only 12 points (2 hours) later than the first degraded operational data appears in the 3th sliding window. If the SCADA data were 1-min averages rather than 10-min averages, the degradation alarm can be generated more quickly.

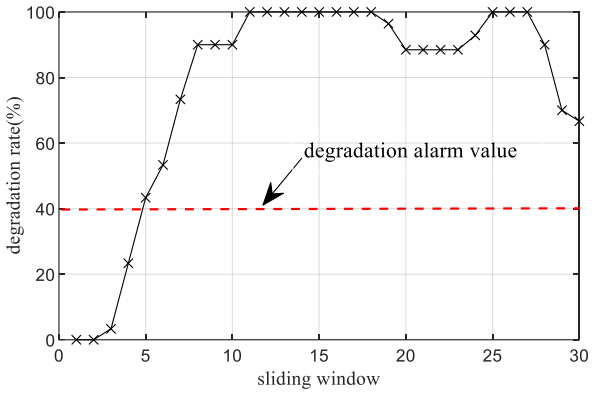
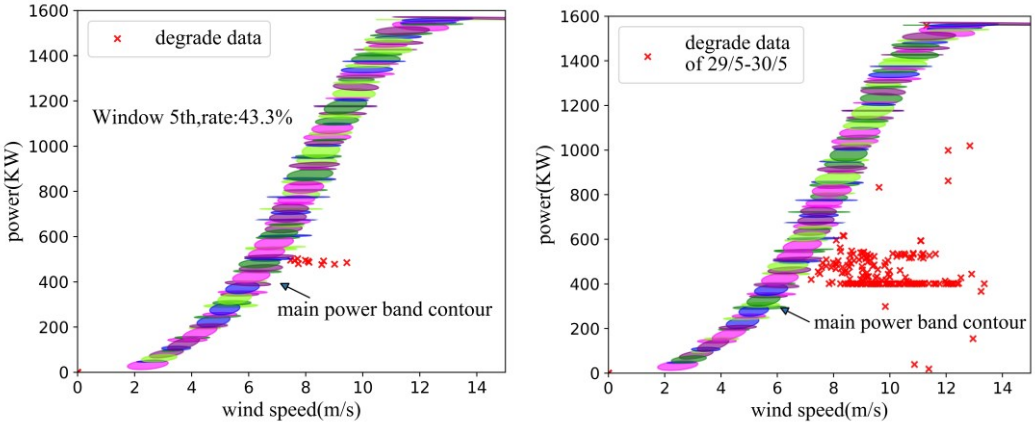


Fig.10 wind turbine degradation monitoring with sliding windows



(a) degrade data in 5th sliding window (b) degrade data of 29/5-30/5

Fig.11 degradation monitoring results with DPGMM

The performance monitoring method in [13] is used as a comparison. [13] created power threshold curve based on the observed variance of data of 1/5 to 28/5. In each

wind bin of 0.5m/s, 2.5 standard deviation of the data is subtracted from the mean power in the bin to build the power threshold curve as Fig.12 shown.

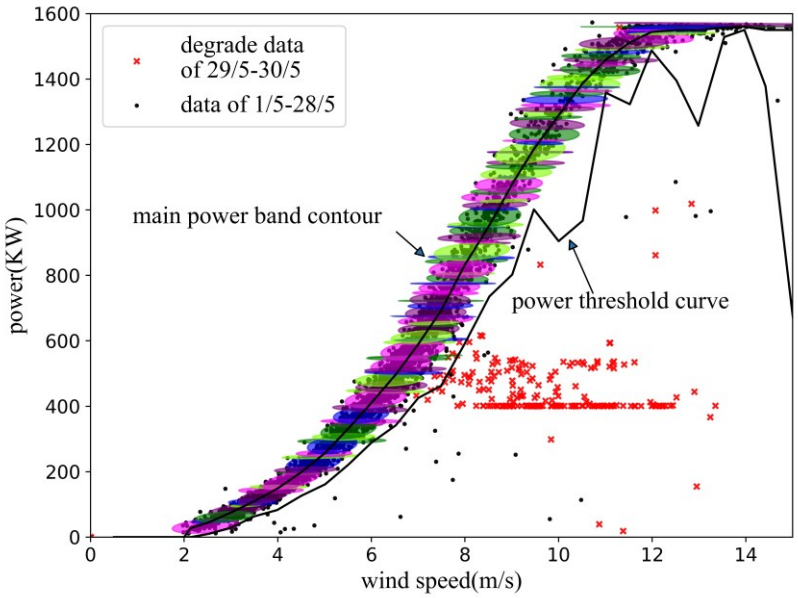
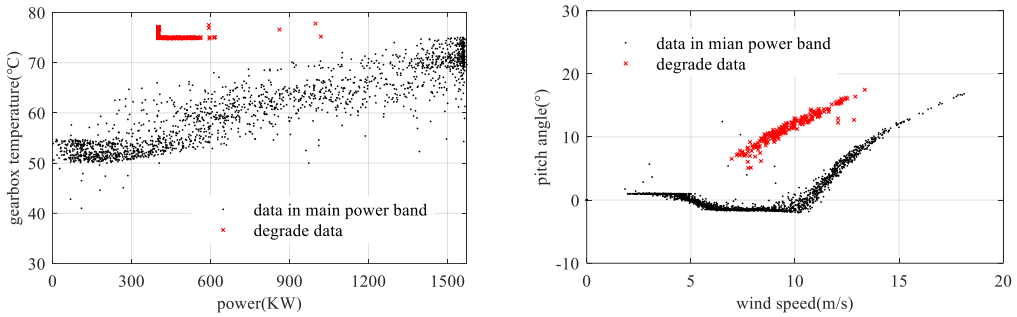


Fig.12 degradation monitoring method comparison

Because there are abnormal data during 1/5 to 28/5, the power variances in some wind bins such as 10-10.5 m/s are large which makes the power threshold curve deviate obviously from the main power band. The method in this paper uses DPGMM to identify abnormal data and the main power band contour correctly depicts the edge of normal data. In Fig.12, we can see some degraded operational data lies outside the main power band contour while within the power threshold curve of [13]. The method in this paper will have a higher degradation monitoring accuracy.

The reason for the performance degradation can be identified from a careful parameter comparison between degraded operational data and the normal data in the main power band for the same wind speed or power working condition as shown in Fig.13. In Fig.13(a), gearbox oil temperatures of degraded operational data during 29/5 to 30/5 reach as high as 75℃ which is much higher than the temperatures of normal data in the main power band at the same power working condition. In Fig.13(b), pitch angles of degraded operational data are also much higher than that of data in the main power band at the same wind speed condition. We infer that during 29/5 to 30/5 the test

wind turbine automatically curtailed output power due to gearbox abnormalities. Because the gearbox oil temperatures reached 75℃, output power was derated with pitch angle obviously larger than for normal operation and wind turbine experienced major performance degradation. The method described in this paper accurately identifies and in a timely manner generates an alarm, in the example only 2 hours after the first signs of degraded operation.



(a) gearbox temperature (b) pitch angle

Fig.13 Gearbox oil temperature and pitch angle for degrade data

5. Conclusions

(1) Normal operational data points in a V-P coordinate system constitute the main power band which is a good reflection of wind turbine generation performance. Data points are partitioned into horizontal power bins. For each power bin DPGMM is used to cluster data points. The envelope contour of the main power band is extracted using confidence ellipses of normal Gaussian components in each power bin and used as the baseline performance model.

(2) An assessment of degraded operational data, based intuitively on Mahalanobis distance, is proposed. Mahalanobis distance is used to measure the distance between the monitoring data and the baseline performance model (the main power band). A degradation degree is defined that can intuitively measure how serious the degradation is for incoming monitoring data.

(3) A sliding window is introduced to improve the reliability of wind turbine performance degradation alarm generation. Whenever the degradation rate exceeds a set value, an alarm is triggered. The reasons for performance degradation can be rooted

out by careful examination of other parameters from the SCADA data. In the example provided, degradation in power output was caused by gearbox oil over temperature, and this confirms the effectiveness of the method.

This paper provides a new approach and method for wind turbine performance degradation monitoring. And the degradation value and degradation alarm value can be treated as adjustable parameters to regulate the sensitivity and reliability of the proposed method in field application.

References

[1] IEC. IEC 61400-12-1: 2005 Power performance measurements of electricity producing wind turbines. 2005.

[2] Yun Wang, Qinghua Hu, Linhao Li, Aoife Foley, Dipti Srinivasan, "Approaches to wind power curve modeling: a review and discussion," Renewable and sustainable energy research, vol.116, pp.109422, Dec.2019.

[3] Helbing Georg, Ritter Matthias, "Improving wind turbine power curve monitoring with standardization," Renewable Energy, vol. 145, pp.1040-1048, Jan.2020.

[4] Davide Astolfi, "Perspectives on SCADA data analysis methods for multivariate wind turbine power curve modeling," Machines, vol.9, no.100, pp.100-117, 2021.

[5] Ayman Quraan, Hussein Masri, Mohammed Mahamodi, Ashraf Radaideh, "Power curve modeling of wind turbine – a comparison study," IET Renewable Power Generation, vol.16, no.2, pp.362-374, Feb.2022.

[6] Peng Guo, David Infield, "Wind turbine power curve modeling and monitoring with Gaussian Process and SPRT," IEEE Transactions on Sustainable Energy, vol.11, no.1, pp.107-115, Jan.2020.

[7] Elena Gonzalez, Bruce Stephen, David Infield, Julio Melero, "Using high-frequency SCADA data for wind turbine performance monitoring: a sensitivity study," Renewable Energy, 2019, vol.131, pp.841-853, Feb.2019.

[8] Yusen He, Andrew Kusiak, "Performance assessment of wind turbines: data-derived quantitative metrics," IEEE Transactions on Sustainable Energy, vol.9, no.1, pp.65-73, Jan.2018.

[9] Huan Long, Long Wang, Zijun Zhang, Zhe Song, Jia Xu, "Data-driven wind turbine power generation performance monitoring," IEEE Transactions on Industrial Electronics, vol.62, no.10, pp.6627-6635, Oct.2015.

[10] Yang Hu, Yunhua Xi, Chenyang Pan, Gengda Li, "Daily condition monitoring of grid-connected wind turbine via high-fidelity power curve and its comprehensive rating," Renewable Energy, vol.146, pp.2095-2111, Feb. 2020.

[11] Xiaodong Jia, Chao Jin, Matt Buzza, Wei Wang, Jay Lee, "Wind turbine performance degradation assessment based on a novel similarity metric for machine performance curves," Renewable Energy, vol.99, pp.1191-1201, Dec.2016.

[12] Silvia Cascianelli, Davide Astolfi, Francesco Castellani, et al, "Wind turbine power curve monitoring based on environmental and operational data," IEEE Transactions on Industiral Informatics, DOI 10.1109/TII.2021.3128205, 2021.

[13] Neil N.Davis, OyvindByrkjedal, Andrea N.Hahmann, "Ice detection on wind turbines using the observed power curve," Wind Energy, vol.19, no.6, pp.999-1010, Jun. 2016.

[14] Peng Guo, David Infield, "Wind turbine blade icing detection with multi-model collaborative monitoring method," Renewable Energy, vol.179, pp.1098-1105, Dec.2021.

[15] Dong Hyuk Yi, Deuk Woo Kim, Cheol Soo Park, "Prior selection method using likelihood confidence region and Dirichlet process Gaussian mixture model for Bayesian inference of building energy models," Energy & Buildings, vol.224, pp.110293, Oct.2020.

[16] Yuelin Li, Elizabeth Schofield, Mithat Gonen, "A tutorial on Dirichlet process mixture modeling," J.Math.Psych, vol.91, pp.128-144, 2019.

[17] Zhenglin Li, Lyudmila Mihaylova, Olga Isupova, Lucile Rossi, "Autonomous flame detection in videos with a Dirichlet process Gaussian mixture color model," IEEE Transactions on Industrial Informatics, vol.14, no.3, pp.1146-1154, Mar.2018.

[18] Yupeng Li, Jianhua Zhang, Zhanyu Ma, Yu Zhang, "Clustering analysis in the wireless propagation channel with a variational Gaussian mixture model," IEEE Transactions on Big Data, vol.6, no.2, pp.223-232, Jun.2020.

[19] Yerebakan, Dundar, "Partially collapsed parallel Gibbs sampler for Dirichlet process mixture models," Patten Recognition Letters, vol.90, pp.22-27, Apr.2017.

[20] Martino Luca, Elvira Victor, Camps Valls, "The recycling Gibbs sampler for efficient learning," Digital Signal Processing, vol.74, pp.1-13, Mar, 2018.

[21] Raul Ruiz, "Wind farm monitoring using Mahalanobis distance and fuzzy clustering," Renewable Energy, vol.123, pp.526-540, Aug.2018.

[22] Rehman Naveed, Khan Bushra, Naveed Khura, "Data-driven multivariate signal denoising using Mahalanobis distance," IEEE Signal Processing Letters, vol.26, no.9, pp.1408-1412, Sep.2019.

[23] Elisa Cabana, Rosa Lillo, Henry Laniado, "Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators," Statistical Papers, vol.62, no.4, pp.1583-1609, Apr.2021.

[24] Peng Guo, David Infield, Xiyun Yang, "Wind turbine generator condition-monitoring using temperature trend analysis," IEEE Transactions on Sustainable Energy, vol.3, no.1, pp.124-133, Jan.2012.