IET Renewable Power Generation

Special Issue: Wind Turbine Condition Monitoring, Diagnosis and Prognosis

Hierarchical method for wind turbine prognosis using SCADA data

ISSN 1752-1416
Received on 31st March 2016
Revised 24th October 2016
Accepted on 25th November 2016
E-First on 18th January 2017
doi: 10.1049/iet-rpg.2016.0247
www.ietdl.org

Niya Chen[1] ✉, Rongrong Yu[1], Yao Chen[2], Hailian Xie[1]
[1]*Power and Control, ABB Corporate Research Center, Beijing, People's Republic of China*
[2]*Electrical Systems, ABB Corporate Research Center, Vasteras, Sweden*
✉ *E-mail: niya.chen@cn.abb.com*

Abstract: Rapid development of wind energy requires effective wind turbine prognosis methods, which can give alarm before actual failure happens and hence enables condition-based maintenance. A hierarchical method based on GP (Gaussian Processes) and PCA (Principal Component Analysis) is proposed in this paper for turbine prognosis using SCADA data. The method includes two levels of prognosis: 1) detect which wind turbine behaves abnormally and has potential defect; 2) determine the defective components in the abnormal turbine. On turbine level, the relationship between selected parameters and power generation is trained based on GP. Then the model residual, which is calculated as the difference between the estimated output and the actually measured power, can indicate whether the turbine is defective. On component level, the contribution of each SCADA variable to turbine abnormality can be given based on PCA method, and can be used for indicating the defective components. Field dataset including 24 failed turbines is used to validate the proposed hierarchical method. The validation results show that the proposed method can achieve wind turbine prognosis with 79% detection rate on turbine level and 76% detection rate on component level. Moreover, the method can provide several months ahead alarm before severe failure happens.

1 Introduction

Utilisation of wind energy has been growing rapidly all over the world [1]. In China, the accumulated capacity of wind farms has reached 145.1 GW, with a growth rate of 26.6% in 2015 [2]. With more and more wind turbines in operation for a long time, wind farm operation and maintenance become more challenging [3]. Failure of key components in wind turbines such as gearbox, generator etc., would lead to a difficult and costly maintenance, and hence bring a catastrophic effect [4]. As a result, there is a growing interest in wind turbine prognosis, which enables condition-based wind turbine maintenance. Compared with scheduled maintenance, condition-based maintenance can reduce the total cost of wind farm operation [5].

Much effort has been put into wind turbine condition monitoring and prognosis [6]. However, most of the investigated methods require installation of additional sensors [7], which is expensive and causes re-certification of wind turbines. Therefore, condition monitoring approach based on supervisory control and data acquisition (SCADA) data is drawing more and more attention these years [8]. SCADA system already exists in wind turbines and can monitor many operating features such as temperature, power generation, blade rotation speed etc. This condition monitoring approach does not need any additional expense and thus can make the overall maintenance more cost-effective.

Different kinds of methodologies have already been applied in SCADA-based wind turbine condition monitoring. Non-linear state estimation technology is used to diagnose gearbox failure in [9]. Neural network and fuzzy logic are combined in [10] for large (2 MW) wind turbine condition monitoring. Three kinds of SCADA-based condition monitoring methods are compared, and the effectiveness of data-mining algorithm in turbine diagnosis is proved in [11]. Performance of multiple wind turbine prognosis methods based on SCADA data are compared in [12], which proves the effectiveness of both artificial neural network and Gaussian processes (GPs) based methods. The comparison in [12] shows that GP-based method has more balanced performance in terms of quite high detection rate and relatively low false alarm rate.

However, only turbine level diagnosis/prognosis or prognosis of single component (e.g. gearbox, blade etc.) was addressed in the prior art research. In this paper, a hierarchical prognosis approach based on SCADA data is proposed, with consideration of both turbine level prognosis and component level prognosis. Four key components are considered in component level prognosis: gearbox, generator, blade, and yaw. The proposed hierarchical prognosis method not only can detect abnormality of wind turbine before failures occur in reality, but also can indicate which component(s) is the root cause of the detected turbine abnormality. Therefore, a better maintenance guidance can be provided. To achieve this functionality, two levels of prognosis are included in the approach: (i) use GP method to detect which wind turbine behaves abnormally and has a defect and (ii) compare the contribution vector obtained by principal component analysis (PCA) method to determine the defective components in the abnormal wind turbine.

GP is a newly developed machine learning algorithm and has been broadly applied in different fields [13, 14]. GP is chosen as the main algorithm to build models in this paper because it has many advantages such as stable performance and low requirement on the number of training samples. More importantly, the performance of GP for wind turbine prognosis has already been proved in [12]. Defective component(s) is identified by PCA method, which calculates a ‘contribution vector’ to indicate the components that have large contribution to turbine abnormality.

Field SCADA data from a large commercial wind farm in northern China is used to train the model and to validate the proposed hierarchical method. The results show that the method can achieve 1–5 months ahead prognosis before severe turbine failure happens, and can also determine defective components with high detection rate and low false alarm rate.

This paper is organised as follows. Section 2 describes the structure of the hierarchical prognosis method including modelling process on both turbine level and component level. Section 3 introduces the methodologies used in the proposed model: GP, automatic relevance determination (ARD), and PCA method. The information of dataset, the steps for data pre-processing, the detailed model validation process, and the experimental results are presented in Section 4. Finally, Section 5 draws the overall conclusions.

2 Hierarchical prognosis approach

2.1 Overall modelling process

Regression model based on GPs has good performance for wind turbine prognosis. However, it is hard to identify defective component in order to provide further maintenance guidance for wind farm operator. Therefore, in hierarchical model, first the SCADA data is inputted into a GP prognosis model to estimate the health condition of wind turbine. Then if this wind turbine is detected as abnormal, a PCA model is used to determine which component(s) is defective, as shown in Fig. 1.

PCA is a classic mathematical method for data dimension reduction. The core idea is to transform the original variables to new variables (principal components) with the target to retain the largest variance in the data and decouple the covariance among different variables. Therefore, the variation in principal components can reflect the condition change of the turbine. A 'contribution vector' can be calculated to evaluate the contribution of each original variable to the variation of new principal components.

The reason for applying PCA method to component level prognosis is that the comparison among the contribution value in contribution vector can indicate the possible defective components: variables with larger contribution value to turbine abnormality means that the corresponding turbine components have higher possibility of being defective.

2.2 Turbine level modelling process

The turbine level prognosis module outputs the turbine status using SCADA data as inputs. There are three main parts in the modelling process of this prognosis module. First, ARD method is applied to select the SCADA variables that are strongly relevant to turbine status as model inputs. The purpose is to reduce the modelling complexity in the next step, since there are dozens of parameters monitored in SCADA system. Second, based on GP, the relationship between the selected input variables and output variable is trained. In this paper, the generated active power is selected as the output variable, because it is a turbine level variable and has the most direct relationship with turbine condition among all SCADA variables. Third, model residual can be calculated as the difference between the estimated output and the actually measured power generation. Then, the calculated residual is used for evaluating turbine condition with a residual larger than predefined threshold indicating a potential failure.

As shown in Fig. 2, first partial SCADA variables that are relevant to turbine condition are selected for modelling. The detailed selection method and results are presented in Section 4.3. Then, the dataset is divided into two parts: training set and test set. The training set will be used for model training and threshold determination, and the test set can be used for model validation. Statuses of all the turbines are known according to the record from wind farm operator; therefore, the performance of prognosis model can be validated by comparing the prognosis result with the status record.

In the training module, normal data is used to build up the model between the selected input variables and the output variable (power generation). This model is supposed to be able to reflect the function between inputs and output correctly when the turbines are normal. After the model is built, it is used to calculate the residual (the difference between the estimation and the measured data) for all the normal data and failure data in the training set. When turbines are normal, the estimated output value shall be close to the measured power generation; conversely, the residual value shall be large for any turbine that is abnormal. Next, a threshold value can be determined to distinguish failure residual and normal residual. This is possible since all the residual data and turbine status are known in the training set. With the determined threshold, the turbine status in the test set can be evaluated. For example, when the residual value is larger than the threshold, the turbine shall be abnormal.

Two outcomes are obtained from the training module: a well-trained model that reflects normal input–output function and a threshold value that can separate abnormal and normal residual. On the basis of these, the current turbine status can be estimated with new input data from the test set. The process is as follows. First, the data of the selected SCADA variables is inputted into the trained normal model and an estimated output will be obtained. Second, the residual can be calculated as the difference between the estimation and the measured data. Finally, the residual is compared with the predefined threshold. If it is smaller than the threshold, the turbine can be determined as normal. Otherwise, there is probably a defect developing in this turbine.

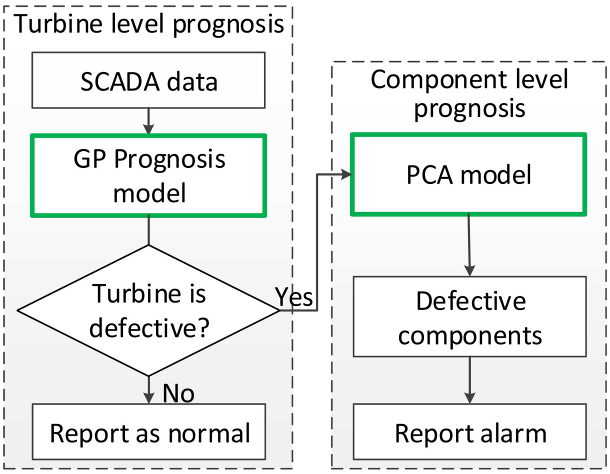

Fig. 1 *Hierarchical modelling process*

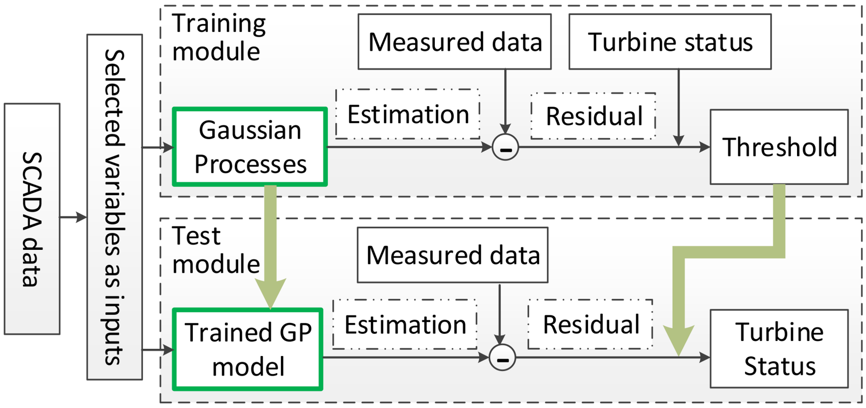

Fig. 2 *Modelling process of turbine level prognosis*

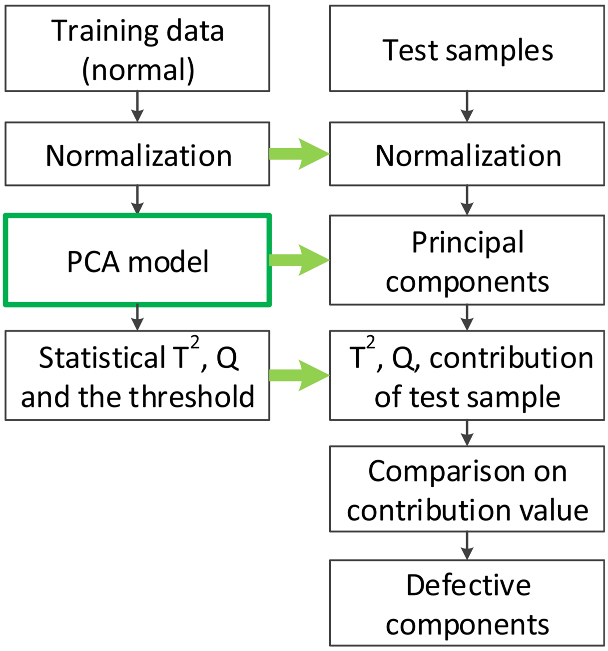

Fig. 3 *Modelling process of component level prognosis*

2.3 Component level modelling process

When a wind turbine is detected as abnormal by the turbine level prognosis process, its real-time SCADA data will be inputted to PCA model for further root cause determination.

As shown in Fig. 3, the training set consists of a certain number of normal samples, which can represent the characteristics of wind turbine in normal condition. With the training set, the following three functions can be obtained, and will be used for testing later: the normalisation function, PCA model for dimension reduction, and the calculation of Hotelling's T^2 and Q statistics and the corresponding threshold [15]. Hotelling's T^2 and Q statistics are both classical statistical indexes to indicate the variation level in data.

The original SCADA data of different variables are in different ranges due to different units. Therefore, they need to be normalised before being putted into the same model together. Then based on PCA model, the normalised variables can be transformed into another lower dimensional space with orthogonal basis. The transformed data can be considered as containing the principal information of the original data and excluding the interaction among different parameters. Therefore, the variation in the transformed data can indicate the health change of the wind turbine. In this paper, the variation is represented by the calculated Hotelling's T^2 and Q statistics.

The real-time SCADA data collected from an abnormal turbine can be considered as a test sample in Fig. 3. After three steps using the three functions obtained in the training phase, the T^2 and Q statistics of the test sample can be calculated accordingly, and the 'contribution vector' of the original SCADA variables can also be obtained. A larger contribution value in the vector means that the corresponding variable has stronger relation to the abnormality appearing in the defective turbine. Mapping of the variables to the turbine components is quite straight forward, which is explained in Section 4. Therefore, by comparing the contribution values, the defective components can be determined as the components related to the variables that have large contribution values. The detailed method and equations of PCA are presented in Section 3.3.

3 Mathematical methods

3.1 Gaussian processes

As introduced in the previous section, the target of GPs is to build up the relationship between the selected input variables and the output variable (power generation). In this paper, the selected variables from SCADA parameters are: ambient temperature, gear oil temperature, gearbox bearing temperature, generator temperature, generator frequency, generator Revolutions per Minute (RPM), blade RPM, yaw angle, and wind speed. The detailed selection process is described in Section 4.3.

The selected input variables are denoted as x and the power generation is denoted as $\boldsymbol{y}$. The training set with n samples can be written as $\{\boldsymbol{D} = f(\boldsymbol{x}_i, y_i) | i = 1, \ldots, n\}$, where $\boldsymbol{x}_i$ is an input vector that includes the data of all the selected input variables collected at time t_i and y_i is a scalar value equal to the measured power generation at the same time point. Collecting the inputs into a matrix $\boldsymbol{X}$ and outputs in a vector $\boldsymbol{y}$, we can write $\boldsymbol{D} = (\boldsymbol{X}, \boldsymbol{y})$. The key point is to model the relationship between the inputs and the output. That is to build a function to satisfy

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i \tag{1}$$

where the observed values $\boldsymbol{y}$ differ from the function values $f(x)$ by additive noise ε_i, which is assumed as following an independent, identically distributed Gaussian distribution with zero mean and variance σ_n^2, i.e. $\varepsilon \sim N(0, \sigma_n^2)$.

Note that $\boldsymbol{y}$ is a linear combination of Gaussian variables and hence is itself Gaussian [16]. Assume the mean function of $\boldsymbol{y}$ to be zero, since the mean function can be easily transformed to zero by subtraction if it is not. The prior on $\boldsymbol{y}$ becomes

$$\begin{aligned} E[y] &= E[f + \epsilon] = 0 \\ \mathrm{cov}[y] &= \mathrm{cov}\left[\boldsymbol{y} \cdot y^{\mathrm{T}}\right] = \mathrm{cov}\left[(f(x) + \epsilon) \cdot (f(x) + \epsilon)^{\mathrm{T}}\right] \\ &= \mathrm{cov}\left[f(x) f(x)^{\mathrm{T}}\right] + \mathrm{cov}\left[\epsilon\epsilon^{\mathrm{T}}\right] = \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 I \end{aligned} \tag{2}$$

where $\boldsymbol{K}$ is a matrix with elements $k_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is the covariance function of x.

Given a training set $\boldsymbol{D} = (\boldsymbol{X}, \boldsymbol{y})$, our goal is to make predictions of power generation based on the test data of the selected input variables, i.e. predict the target variable y_* based on a new input $\boldsymbol{x}_*$. Since we already have $p(\boldsymbol{y} | \boldsymbol{X}, k) = N(0, \boldsymbol{K} + \sigma_n^2 I)$, the distribution with new input can be written as

$$\begin{bmatrix} y \\ y_* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 I & k(\boldsymbol{X}, \boldsymbol{x}_*) \\ k(\boldsymbol{x}_*, \boldsymbol{X}) & k(\boldsymbol{x}_*, \boldsymbol{x}_*) \end{bmatrix}\right) \tag{3}$$

where $k(\boldsymbol{X}, \boldsymbol{x}_*) = k(\boldsymbol{x}_*, \boldsymbol{X})^{\mathrm{T}} = [k(\boldsymbol{x}_1, \boldsymbol{x}_*), \ldots, k(\boldsymbol{x}_n, \boldsymbol{x}_*)]^{\mathrm{T}}$ and will be abbreviated as k_*. Then according to the principle of joint Gaussian distributions, the prediction result for the target is given by

$$\begin{aligned} \overline{y_*} &= k_*^{\mathrm{T}} (\boldsymbol{K} + \sigma_n^2 I)^{-1} \boldsymbol{y} \\ V[y_*] &= k(\boldsymbol{x}_*, \boldsymbol{x}_*) - k_*^{\mathrm{T}} \left(\boldsymbol{K} + \sigma_n^2 I\right)^{-1} k_* \end{aligned} \tag{4}$$

Note that the equation above is completely explicit since parameter values in covariance function k can be trained by training set $\boldsymbol{D}$. Now, the whole regression model based on GP is completed, with the expectation $\overline{y_*}$ as the predicted value of power generation based on test data $\boldsymbol{x}_*$, and $V[y_*]$ as the confidence interval of prediction. Finally, the residual error of the test data $(\boldsymbol{x}_*, y_*)$ can be calculated as follows:

$$R_* = y_* - \overline{y_*} \tag{5}$$

When the turbine is in normal state, the GP model should fit well. Therefore, the predicted value is very close to the actually measured power, which leads to a residual error close to zero.

3.2 Automatic relevance determination

The first step in turbine level prognosis is to select SCADA variables that are relevant to turbine condition as model inputs. The selection process is based on ARD method, which is a part of GPs theory.

In practise, the parameter values of covariance function in GP model can be inferred from the observed training data. This learning process, which is usually called 'learning the hyperparameters', can be accomplished by maximising the log likelihood function [16]

$$\ln p(\boldsymbol{y} \mid \theta) = -\frac{1}{2}\ln|\boldsymbol{K}| - \frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}^{-1}\boldsymbol{y} - \frac{n}{2}\ln\pi \quad (6)$$

This function can be maximised using a standard non-linear gradient-based optimisation algorithm. We can extend this technique by incorporating a separate parameter for each input variable. Then, the relative importance of different inputs can be inferred from the observed data. This leads to the ARD method, which can be used to choose the inputs that have a significant influence on the outputs. For example, assume that the covariance function is the squared exponential and the original input vector $\boldsymbol{x}_i$ has n dimensions

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \theta_0 \exp\left\{-\frac{1}{2}\sum_{k=1}^{n} l_k (x_i^k - x_j^k)^2\right\} + \sigma_n^2 \delta \quad (7)$$

where $\{\theta_0, l_1, l_2, \ldots, l_n\}$ are the 'hyperparameters' that can be learned based on the training set. All the length scale parameters can be contained in a vector $\boldsymbol{L} = \{l_1, \ldots, l_n\}$. The hyperparameters $\boldsymbol{L}$ are used to implement ARD, since the value of $\boldsymbol{L}$ obviously indicates how relevant the input is: as l_k becomes larger, the function becomes more sensitive to the corresponding input variable. Therefore, the importance of each input variable is revealed, and the inputs with small value of l_k can be discarded. For example, if SCADA variable 'ambient temperature' is denoted as x^k, and the corresponding length scale l_k is very small, this variable should not be selected as an input to the GP model. The ARD toolbox in Netlab [17] is used to determine the appropriate inputs for GP model in this paper.

3.3 Principal component analysis

When a wind turbine is alarmed as abnormal by turbine level prognosis module, PCA is used to determine which component(s) is the cause of turbine abnormality. PCA is a linear unsupervised optimal dimension reduction technique in terms of capturing the variance of data [18]. Given a data matrix $[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m]$ with m observations of n variables

$$\boldsymbol{D} = \begin{bmatrix} x_1^1 & \cdots & x_m^1 \\ \vdots & \ddots & \vdots \\ x_1^n & \cdots & x_m^n \end{bmatrix} \quad (8)$$

In matrix $\boldsymbol{D}$, each row vector $\boldsymbol{x}^j$ contains all the data of one SCADA variable, and each column vector $\boldsymbol{x}_t$ is the data recorded for all variables at one time point. PCA can reduce the dimension n by transforming matrix $\boldsymbol{X}$ into a new matrix $\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{X}$ in a space with orthogonal basis $[p_1, p_2, \ldots, p_k]$, where the new variables are uncorrelated to each other. The transformation is defined as

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{p}_1^{\mathrm{T}} \\ \boldsymbol{p}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{p}_k^{\mathrm{T}} \end{bmatrix} * [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m] \quad (9)$$

The transformation has two targets. First is that the new variables should capture most variance of the data. Second is that the covariance among new variables should be decreased to the minimum value. This means the covariance matrix $\boldsymbol{C}$ of new variables should be a diagonal matrix, which can also be interpreted as follows:

$$\boldsymbol{C} = \boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}} = \boldsymbol{P}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})P^{\mathrm{T}} \quad (10)$$

Therefore, it is obvious that the proper matrix $\boldsymbol{P}$ is the one that can diagonalise the covariance matrix of $\boldsymbol{X}$. According to classical matrix theory, this matrix $\boldsymbol{P}$ can easily be calculated as follows. The orthogonal eigenvector $\boldsymbol{E} = [e_1, e_2, \ldots, e_n]$ of the $n \times n$ real symmetric matrix $\boldsymbol{M} = \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ can surely be found

$$E^{\mathrm{T}}\boldsymbol{M}\boldsymbol{E} = E^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})\boldsymbol{E} = \begin{pmatrix} \lambda_1 & 0\cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0\cdots & \lambda_n \end{pmatrix} \quad (11)$$

Then, sort the eigenvalues $[\lambda_1, \lambda_2, \ldots, \lambda_n]$ in a descending order. The vectors corresponding to the largest k eigenvalues can be selected as principal components, which compose matrix $\boldsymbol{P} = [e_1, \ldots, e_k]^{\mathrm{T}}$.

After the dimension reduction, given a new observation $\boldsymbol{x}_i$, the variation in $\boldsymbol{y}_i = \boldsymbol{P}\boldsymbol{x}_i$ can be calculated in both the score space and the residual space. Hotelling's T^2 statistic evaluates the conformity of the new observation with the PCA model in the principal component space while Q statistic evaluates the conformity from the perspective of residual. Assume that each variable follows Gaussian distribution, the distributions of T^2 and Q can be approximated and the confidence limits can be derived [15].

More importantly, for component level prognosis, the 'contribution vector' of SCADA variables can be plotted. Contribution plot shows the contribution of each variable to statistics and is very popular in PCA to fulfil fault diagnosis [19]. For example, given a test vector $\boldsymbol{x} = [x^1, \ldots, x^j, \ldots, x^n]^{\mathrm{T}}$, the contribution to Hotelling's T^2 statistic of the variable x^j can be calculated as follows:

$$Con\_T_j = \sum_{i=1}^{k} \mathrm{cont}_{i,j} = \sum_{i=1}^{k}\left(x^j p_{j,i} \frac{\boldsymbol{p}_i^{\mathrm{T}} * \boldsymbol{x}}{\lambda_i}\right) \quad (12)$$

where j indicates the variable number and $p_{j,i}$ is the jth variable in p_i.

In general, variables contributing much higher than the others are regarded as responsible for the detected faults. In this paper, a loose rule is set to define 'higher contribution' due to the specialty of component level prognosis for wind turbine: the cost of examining one more component is much lower than missing one defective component once the wind turbine is shut down for maintenance.

The detailed rule is: parameters that have contribution larger than half of the largest contribution are considered as parameters reflecting abnormality

$$F_p = \left\{x^j \mid Con\_Tj > \max\left(\overrightarrow{Con\_T}\right)\right\} \quad (13)$$

The comparison based on Q statistic is similar to the equation on T^2 statistic.

4 Experimental validation

4.1 Dataset

A field dataset from a large wind farm (one of the top largest wind farms in China) was used in this paper to evaluate the proposed hierarchical prognosis approach. The wind farm (denoted as Farm-T) is located in the northern part of China, where rich wind resource and volatile weather raise strong demand for turbine prognosis/diagnosis system. General information of Farm-T is given in Table 1.

As shown in Table 1, Farm-T consists of double fed induction generator (DFIG) type wind turbines with a nominal power of 850 kW. DFIG is the most popular wind turbine type in onshore wind farms. Those 850 kW DFIGs are popular wind turbines installed several years ago, and thus need diagnosis and maintenance now. Therefore, the data chosen for verification can represent typical cases.

The dataset includes SCADA data and failure reports gathered from Farm-T during the period from March 2012 to March 2014, when failures occurred to 26 wind turbines. In the failure report, which is provided by wind farm operator, not only the turbine status is recorded, but also the defective components are recorded when a turbine failure happens. There are four kinds of components recorded in the failure reports: gearbox, pitch, generator, and yaw.

The dataset also contains data from wind turbines that were constantly normal. For model validation, data from two failed turbines and three normal turbines are collected in the training set and the data from the rest 24 failed turbines are gathered in the test set.

As mentioned earlier, only SCADA data are used as model inputs in the proposed prognosis method, which makes the prognosis method cost-effective. The SCADA system in Farm-T includes 24 parameters, as listed in Table 2.

It can be seen from Table 2 that not all the parameters would be useful for wind turbine prognosis/diagnosis. In fact, many parameters are obviously not related to turbine condition, e.g. statistical value 'total power generation'. Therefore, parameters that actually influence the condition of turbine are chosen in Section 4.3.

4.2 Data pre-processing

Original SCADA data from wind farm includes many error data caused by communication problem or sensor failure. Therefore, pre-processing steps are needed to avoid influence of wrong data:

i. *Mark error points with communication problem:* When there is a communication problem in the SCADA system, the received data will be zero. Therefore, the rows in the data matrix with all zeroes are marked, since some parameter should not be zero at any time point, such as 'total power generation'.
ii. *Mark unreasonable data:* Failed sensor can send back obvious unreasonable data, such as wind speed higher than 1000 m/s, wind direction larger than 360° etc. All the rows in data matrix with unreasonable values are marked.
iii. *Mark unreasonable peak points:* Occasional disturbances on sensor might lead to peak values in measurement. Peak value means one extremely high or low value compared with the previous and the next measurement, which is impossible due to physical principle.
iv. Count the number of marked points of each parameter, and the variables that with >25% proportion of marked points are discarded. That is because wrong results might be obtained from variables with too many wrong information recorded.
v. Apply cubic spline interpolation method [20] to create more reasonable data for all the marked samples in the variables left.

Table 1 Information of Farm-T

Turbine type	Rated power	Data duration	Failure number	Failed components
DFIG[a]	850 kW	2012.3–2014.3	26 turbines	gearbox, pitch, generator, yaw

[a]Double fed induction generator.

Table 2 Parameters in SCADA

Ambient temperature	A/B/C phase current	generator frequency	total power generation
Gear oil temperature	A/B/C phase voltage	generator RPM	total effective time
Gearbox bearing temperature	AC power	blade RPM	data time
Generator temperature	reactive power	yaw angle	fault time
Nacelle temperature	power factor	wind direction	wind speed

Table 3 ARD results of Farm-T

Ambient temperature	Gear oil temperature	Gearbox bearing temperature	Generator temperature
0.17	3.39	0.17	9.39
Generator frequency	Generator RPM	Reactive power	Nacelle temperature
9.69	2.44	0.0001	0.001
Blade RPM	Yaw angle	Wind direction	Wind speed
0.06	5.50	1.67	0.79

4.3 Parameter selection

To avoid waste of effort and reduce model complexity, parameters in SCADA that are actually related to turbine condition are chosen in this part.

First, 11 parameters can be discarded by manual selection: (i) obviously irrelevant parameters that are recorded just for review: total power generation, total effective time, data time, and fault time; (ii) redundant parameter: power factor; and (iii) discarded parameters in previous pre-processing step: A/B/C phase current and phase voltage.

Second, since 'AC power' indicates the real-time power generation of the wind turbine, it can directly reflect the overall turbine condition. Hence it is chosen as 'turbine condition indicator', i.e. the output of turbine level prognosis model. ARD method is applied to determine which SCADA variables should be included as inputs to the prognosis model. Using the 'AC power' as the target variable, the relevance values are listed in Table 3.

It can be noted that some variables have relatively small relevance values. According to Table 3 and considering the trade-off between the number of retained useful parameters and model complexity, the following variables are finally chosen as model inputs: ambient temperature, gear oil temperature, gearbox bearing temperature, generator temperature, generator frequency, generator RPM, blade RPM, yaw angle, and wind speed.

4.4 Case study

To avoid the influence of different scales of variables, the data of all the selected inputs and output variables are normalised, using the classic normalisation equation

$$\boldsymbol{y} = (y_{\max} - y_{\min}) * (\boldsymbol{x} - x_{\min})/(x_{\max} - x_{\min}) + y_{\min} \tag{14}$$

where $y_{\max} = 1$ and $y_{\min} = 0$. After normalisation, a GPs based model is trained by part of the normal data in the training set. Then, the residuals corresponding to other data in the training set can be obtained by calculating the difference between model output and the measured power generation. The residual curve of turbine #25 is presented in Figs. 4 as an example.

Note that the condition of turbine #25 is already known and is shown in Fig. 4: the data from the first 25 days are indicated as normal data, and there is a gearbox failure happening on the 88th day. It is clearly reflected in this figure that the residual error stays in a quite small range while wind turbine is in normal condition, and increases with oscillations as time approaches the failure point.

A reasonable residual threshold should separate the normal and abnormal data well, thus provide an advanced alarm for gearbox failure, i.e. prognosis. On the basis of the analysis on the residuals of all the normal samples in the training set, the threshold is set by the following equation:

$$t = \begin{array}{l} \max\,(\mathrm{mean}(\mathrm{residual}_{\mathrm{trainNormal}}) \times 3, \\ \max\,(\mathrm{residual}_{\mathrm{trainNormal}}) \times 2) \end{array} \tag{15}$$

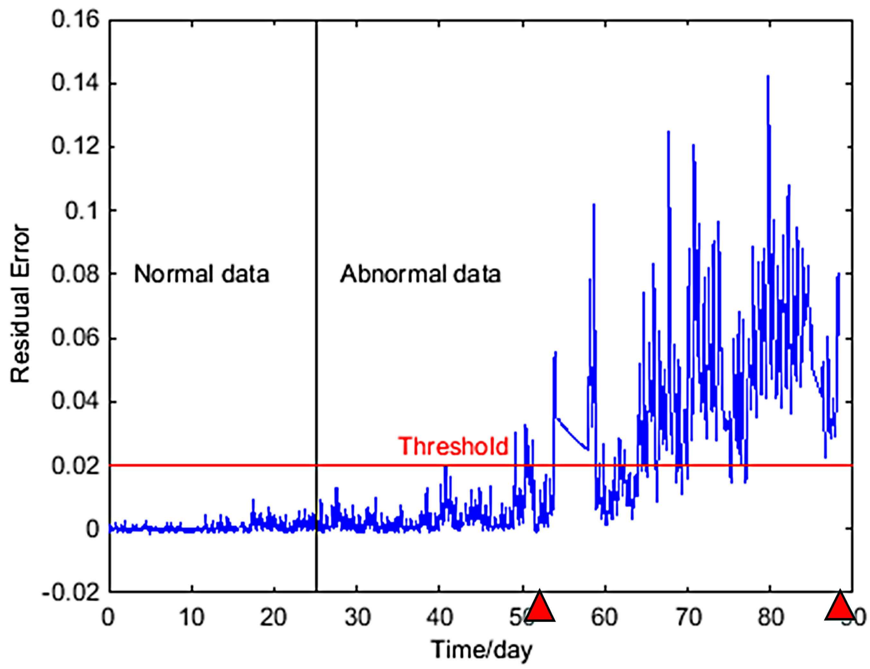

Fig. 4 *Residual error of turbine #25 by trained GP model*

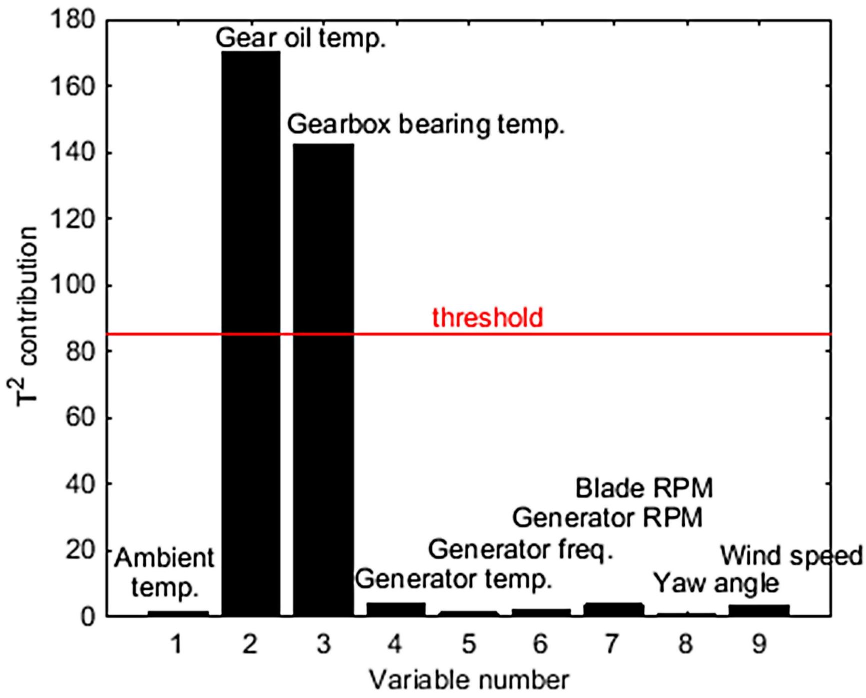

Fig. 5 *Contribution plot of turbine #25 by Hotelling T^2 statistic*

where $\text{residual}_{\text{trainNormal}}$ is the series of residual errors of the normal samples in the training set.

As soon as the turbine level prognosis model reports an alarm, component level model is called to determine the defective component list. Let us still take turbine #25, for example, the contribution vector can be plotted since an alarm is created around the 52nd day. Test samples are selected from SCADA data recorded in 3 days after the alarm point, in order to calculate the average contribution value of each parameter in the 3 days. The contribution vector on T^2 statistic is plotted in Fig. 5.

It can be clearly seen from this figure that gearbox-related variables have higher contribution. Therefore, the component level model finally reports to the wind farm operator that the gearbox is the defective component – which is correct according to the failure report. The 3 days data after alarm point are used in order to eliminate random factors. That also means component level result can be several days later than turbine level alarm.

4.5 Experimental results

To evaluate the performance of the developed model, two performance indexes are used in this paper: detection rate and false alarm times. The definitions are as follows:

$$\boldsymbol{D} = n_{\text{detected}}/N_{\text{number of WTs}} \times 100\% \tag{16}$$

$$F = n_{\text{false alarm}}/N_{\text{number of WTs}}/t_{\text{year}} \tag{17}$$

Detection rate $d\%$ means out of 100 failed turbines, d turbines are successfully detected as abnormal before the real failure comes. False alarm times F/turbine/year means that for one normal turbine/component in 1 year, it is wrongly alarmed as abnormal for F times.

Data from 24 wind turbines having failures reported are collected in the test set. After applying the hierarchical prognosis

Table 4 Experimental results of hierarchical prognosis method

Failed wind turbine (WT) no.	Estimated turbine status	Prognosis time	Actual defective components	Estimated defective components (by T^2)	Estimated defective components (by Q)
WT #1	failed	3 months	gearbox, generator	generator, blade, yaw	gearbox, generator, blade
WT #2	failed	1 month	gearbox, generator	generator	gearbox, generator, blade
WT #3	normal	—	—	—	—
WT #4	failed	2 months	generator	generator, gearbox, yaw	generator
WT #5	failed	3 months	generator, gearbox	generator, gearbox, blade	generator, gearbox, blade
WT #6	normal	—	—	—	—
WT #7	failed	3 months	generator	generator, gearbox, blade	generator, blade
WT #8	failed	4 months	generator	yaw	generator
WT #9	failed	4 months	gearbox	generator, blade, yaw	yaw
WT #10	failed	1 month	generator	gearbox, generator	yaw, blade
WT #11	normal	—	—	—	—
WT #12	failed	2 months	generator	gearbox, generator, blade	generator, blade
WT #13	normal	—	—	—	—
WT #14	failed	5 months	generator, gearbox	generator	generator
WT #15	failed	6 months	gearbox	gearbox, generator, blade	generator, blade
WT #16	failed	1 month	gearbox	gearbox	gearbox, blade
WT #17	failed	4 months	gearbox	generator, blade	generator
WT #18	failed	0.5 months	blade	gearbox, generator	generator, blade
WT #19	failed	2 months	generator	generator	generator
WT #20	failed	2 months	generator	generator	generator
WT #21	failed	2 months	yaw, blade	blade	blade
WT #22	normal	—	—	—	—
WT #23	failed	1 month	generator	generator	generator
WT #24	failed	1 month	gearbox, generator	gearbox, generator, blade	gearbox, generator, blade
statistics	turbine level detection rate: 19/24 = 79.2%	average prognosis time: 47.5/19 = 2.5 months	25 component failures	detection rate: 17/25 = 68% false alarm: 21/24/2 = 0.44	detection rate: 19/25 = 76% false alarm: 13/24/2 = 0.27

process to the whole set, the experimental results are listed in Table 4.

It can be seen from Table 4 that 19 out of the 24 turbines in the test set are correctly detected as abnormal, which gives a detection rate of 79.2% on turbine level. Furthermore, the method achieves 2.5 months (on average) ahead prognosis. This achievement shall be very useful for wind farm operation and maintenance. Meanwhile the turbine level false alarm is zero, since no false alarm is raised during the period when the turbines are recorded as normal. On component level, the performance varies with different statistics: the detection rate is 68%, and false alarm is 0.44 times/turbine/year based on T^2 statistic; meanwhile, the detection rate is 76% and false alarm is 0.27 times/turbine/year using Q statistic. A possible explanation on better performance of using Q statistic is that the working condition of wind turbine changes a lot, which naturally influences the T^2 statistic.

5 Conclusion

With the rapid wind power development, the demand of cost-effective wind farm condition monitoring technology is growing. A hierarchical prognosis approach that uses only SCADA data is proposed in this paper for wind turbine condition monitoring. On turbine level, GP regression model is trained to give alarm before real failure happens; on component level, the defective component is determined with PCA model by comparing the contribution value of each parameter to turbine abnormality.

Field dataset from a wind farm in China is used to train the model and to validate the proposed method. Data from two failed wind turbines and three normal turbines are collected in the training set, and data from 24 failed turbines are gathered in the test set.

The experimental results show that 19 out of 24 turbines in the test set are correctly reported as abnormal several months ahead of the real failure occurrence, which means the accuracy of the proposed method on turbine level is 79.2%. Meanwhile there is no false alarm. On component level, the performance using Q statistic is validated as having a 76% detection rate and 0.27 times/turbine/year false alarm rate.

6 Acknowledgment

The authors thank Dr. Shen for providing the valuable wind farm operation data.

7 References

[1] Kaldellis, J.K., Zafirakis, D.: 'The wind energy evolution: a short review of a long history', *Renew. Energy*, 2011, **36**, (7), pp. 1887–1901

[2] Council, G.W.E.: 'Global wind statistics 2015'. Global Wind Report, 2016, pp. 1–4

[3] Milborrow, D.: 'Operation and maintenance costs compared and revealed', *Wind Stats 2006*, 2006, **19**, (3), p. 3

[4] Ribrant, J., Bertling, L.: 'Survey of failures in wind power systems with focus on Swedish wind power plants during 1997–2005'. IEEE Power Engineering Society General Meeting, June 2007

[5] Márquez, F.P.G., Tobias, A.M., Pérez, J.M.P., *et al.*: 'Condition monitoring of wind turbines: techniques and methods', *Renew. Energy*, 2012, **46**, (10), pp. 169–178

[6] Daneshi-Far, Z., Capolino, G.A., Henao, H.: 'Review of failures and condition monitoring in wind turbine generators'. 2010 XIX Int. Conf. on Electrical Machines (ICEM), September 2010

[7] Heming, L., Shuting, W., Yonggang, L., *et al.*: 'Condition monitoring of generator stator winding inter-turn short circuit fault based on electrically excited vibration'. 2005 IEEE Int. Conf. on Electric Machines and Drives, May 2005

[8] Kim, K., Parthasarathy, G., Uluyol, O., *et al.*: 'Use of SCADA data for failure detection in wind turbines'. 2011 Energy Sustainability Conf. and Fuel Cell Conf., August 2011

[9] Wang, Y., Infield, D.: 'Supervisory control and data acquisition data-based non-linear state estimation technique for wind turbine gearbox condition monitoring', *IET Renew. Power Gener.*, 2013, **7**, (4), pp. 350–358

[10] Schlechtingen, M., Santos, I.F., Achiche, S.: 'Wind turbine condition monitoring based on SCADA data using normal behavior models part 1: system description', *Appl. Soft Comput.*, 2013, **13**, (1), pp. 259–270

[11] Wilkinson, M., Darnell, B., van Delft, T., *et al.*: 'Comparison of methods for wind turbine condition monitoring with SCADA data'. EWEA 2013 Annual Event, February 2013

[12] Chen, N., Yu, R.R., Chen, Y.: 'Comparison on wind turbine condition monitoring methods using SCADA data'. Wind Integration Workshop, 2015, no. 025

[13] Salimi-Khorshidi, G., Nichols, T.E., Smith, S.M., *et al.*: 'Using Gaussian-process regression for meta-analytic neuroimaging inference based on sparse observations', *IEEE Trans. Med. Imag.*, 2011, **30**, (7), pp. 1401–1416

[14] Bazi, Y., Melgani, F.: 'Gaussian process approach to remote sensing image classification', *IEEE Trans. Geosci. Remote Sens.*, 2010, **48**, (1), pp. 186–197

[15] Joe Qin, S.: 'Statistical process monitoring: basics and beyond', *J. Chemometr.*, 2003, **17**, pp. 480–502

[16] Rasmussen, C.E., Williams, C.K.I.: '*Gaussian processes for machine learning*' (MIT Press, UK, 2006), pp. 7–32

[17] Nabney, I.T.: '*Netlab: algorithms for pattern recognition*' (Springer, UK, 2004), pp. 369–380

[18] Jolliffe, I.T.: 'Principal component analysis', *Series: Springer Series in Statistics* (Springer, NY, 2002, 2nd edn.), vol. **XXIX**, no. 487, p. 28 illus

[19] Westerhuis, J.A., Kourti, T., MacGregor, J.F.: 'Comparing alternative approaches for multivariate statistical analysis of batch process data', *J. Chemometr.*, 1999, **13**, pp. 397–413

[20] Subbotin, Y.N.: 'Interpolating splines', in Cieselski, Z., Musielak, J. (EDs.): '*Approximation theory*' (Reidel, Netherlands, 1975) pp. 221–234