energies

MDPI

Article

Simultaneous Fault Detection and Sensor Selection for Condition Monitoring of Wind Turbines

Wenna Zhang [1,2,*] and Xiandong Ma [2]

[1] College of Mechatronics and Automation, National University of Defense Technology, Changsha 410073, China

[2] Engineering Department, Lancaster University, Bailrigg, Lancaster LA1 4YW, UK; xiandong.ma@lancaster.ac.uk

* Correspondence: zwna@nudt.edu.cn; Tel.: +86-731-8457-5321

Academic Editor: Frede Blaabjerg

Received: 1 February 2016; Accepted: 7 April 2016; Published: 12 April 2016

Abstract: Data collected from the supervisory control and data acquisition (SCADA) system are used widely in wind farms to obtain operation and performance information about wind turbines. The paper presents a three-way model by means of parallel factor analysis (PARAFAC) for wind turbine fault detection and sensor selection, and evaluates the method with SCADA data obtained from an operational farm. The main characteristic of this new approach is that it can be used to simultaneously explore measurement sample profiles and sensors profiles to avoid discarding potentially relevant information for feature extraction. With *K*-means clustering method, the measurement data indicating normal, fault and alarm conditions of the wind turbines can be identified, and the sensor array can be optimised for effective condition monitoring.

Keywords: wind turbines; supervisory control and data acquisition (SCADA) data; parallel factor analysis (PARAFAC); *K*-means clustering; condition monitoring

1. Introduction

Nowadays, wind power is considered as one of the most viable and sustainable resources worldwide [1]. Wind turbines often operate offshore in order to take advantage of stronger and more reliable winds; however, unscheduled maintenance due to unexpected failures can be costly, not only for maintenance support but also due to lost production time [2]. Condition monitoring systems (CMS) can play a pivotal role in establishing a condition-based maintenance and asset management. Most modern wind turbines incorporate on-board supervisory control and data acquisition (SCADA) systems for control and monitoring of turbine operation and performance. A SCADA system may contain massive amounts of data related to hundreds of parameters of the wind turbines, which has attracted great research interests in fault diagnosis and prognosis for wind turbines.

Usually, one or more parameters generated by SCADA systems are selected and used to obtain models of the turbines operating under different conditions. There are two types of modelling methods: mechanistic methods and data driven model-based methods. The former require a thorough understanding of the process and may result in complex models. The latter do not require knowledge of the process or specific parameters; they are obtained directly from measured input and output signals [3]. For example, Marvuglia [4] investigated an artificial neural network (ANN) model of the relationship between wind speed and generated power for an entire wind farm. The power curves modelled in this way are used to detect faults of the wind farm as a whole. Philip *et al.* [5] proposed a multiple-input (wind speed and active power) single-output (gearbox bearing temperature) state dependent parameter (SDP) method; multivariate SDP models were used to identify the distinct warning levels of a developing fault using adaptive thresholds.

At the same time, in order for the SCADA system to work more accurately, it is essential to obtain enough information about the wind turbine's operational condition and performance. This can be done by using different types of sensors and by monitoring different locations within the wind turbine. Due to the complexity of the turbine, there could be more than 250 monitoring points required to monitor most subsystems of a turbine; the number of the monitoring points will thus be considerably larger for a wind farm [6]. Apart from the large amount of data needing to be handled and transmitted, other questions may have also risen; for example, concerning the redundancy among monitoring data, many of the signals might be highly correlated.

This paper therefore proposes a novel model built using parallel factor analysis (PARAFAC) for fault detection and sensor selection of wind turbines based on SCADA data. As with other related decomposition methods, such as Tucker3 [7] and unfolded principal component analysis (PCA) [8], PARAFAC belongs to the same family of bi-linear or multi-linear methods of decomposing multi-way data into a set of loading and score matrices. However, PARAFAC uses less degrees of freedom than Tucker3 or unfolded PCA methods. This intrinsic feature leads to simpler models and avoids the incorporation of non-significant effects such as noise and redundant information in the model. Originated from psychometrics [9], PARAFAC has attracted increasing interest because it is a processing technique capable of simultaneously determining the pure contributions to the dataset and optimising each factor at a time in trilinear systems. Therefore it has been used in psychology, chemometrics and other areas [10,11]. One of the most popular applications is modelling fluorescence excitation-emission data, which is a commonly used data type in chemistry, medicine and food science. Several studies have been done to explore the underlying chemical phenomena in fluorescence spectral data obtained from sugar solutions in order to investigate quality issues [12] and a fish dataset with known fluorophores [13]. The main characteristic of this new approach is that it can simultaneously explore information regarding sensor contribution and measurement data contribution at different points. Based on such information, by using an appropriate clustering method, measurement samples can be classified and the sensor array can be optimised. However, this method has not previously been applied to condition monitoring of wind turbines.

This paper is organised as follows: wind turbine data used to evaluate the proposed method are presented and pre-processed in Section 2. Section 3 proposes the methodology of the PARAFAC model while Section 4 presents the *K*-means clustering method used to classify measurement data into alarm events, normal and faulty segments. In Section 5 the models are applied to SCADA data obtained from one of the operational wind turbines where the results are cross checked to ensure real faults have been identified. In Section 6 conclusions are drawn and suggestions made for future work.

2. Wind Turbine Data

The SCADA data used for this research were obtained from an operational wind farm. For each turbine, a complete history of sensor information and turbine status information for a period of 16 months are available. These data, with a cut-in wind speed of 3 m/s and a cut-out wind speed of 25 m/s, consist of 128 parameters for various temperatures and pressures, power outputs, vibrations, wind speed, digital control signals and others, associated with the condition parameters of blades, nacelle, rotor, generator, gearbox, grid, hydraulic fluid, cooling water, and meteorological conditions. The SCADA system acquires data at a sample rate in the order of 2 s. The data are then processed and stored at 10-minute intervals in order to significantly reduce the amount of data that need to be processed while still reflecting the operation of wind turbines under normal and fault conditions. Thus a total of 77241 measurements are obtained for each parameter for the period of 16 months. The SCADA data from one of the operational wind turbines are selected for analysis and validation of the proposed method. Alarm logs that record the time at which the alarm occurs and the message that reveals the malfunction of particular parameters of the turbine are also available, which are used to cross-check potential faults identified from the data against what was actually happening.

2.1. Data Selection

Data selection is first carried out to eliminate those digital and constant signals, which are ineffective to the PARAFAC analysis. The meteorological parameters such as wind direction, humidity, air pressure, and those parameters representing set points and digital signals from controllers, together with those parameters that remain constant, are removed from the SCADA data. As one of the most important influencing meteorological parameters, the wind speed is still retained, but it is not used for PARAFAC. Thus there are 52 sensor signals left for PARAFAC analysis, which are associated with the parameters defining the performance of the turbine operations, such as the nacelle position (sensor 1); blades positions (sensors 2–8); mains currents (sensors 9–11); apparent power and active power (sensors 12–13); reactive power (sensor 14); pitch motor currents (sensors 15–17); oil pressures (sensors 18–19); oscillation signals (sensors 20–25), speeds of the generator and the rotor (sensors 26–32), temperatures of the generator windings, the gearbox bearings, the nacelle, the gearbox oil sump, the hydraulic fluid and the cooling water (sensors 33–52).

2.2. Data Pre-Processing

Gaps in SCADA data exist due to occasions when the turbine is inactive during periods of low and high wind speeds. Additional gaps occur due to the occurrence of scheduled maintenance and faults. Prior to the PARAFAC model analysis, it is necessary to remove these gaps. Thus, 45,654 measurement points remained for each turbine parameter. After removal of all the gaps in the data, the active power of the turbine is plotted in Figure 1 as an example.

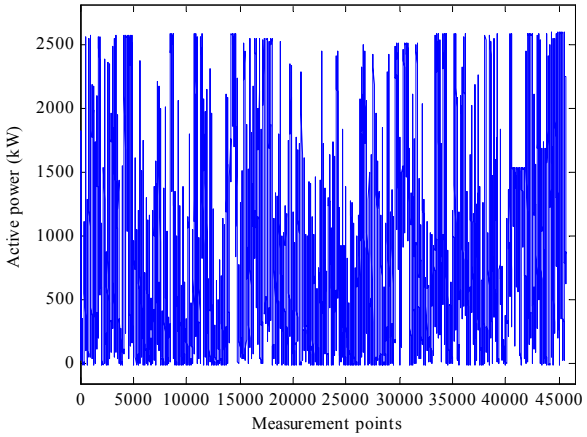


Figure 1. Active power of the turbine after removal of gaps.

Figure 2a,b show the relationship between wind speed and active power output, and between wind speed and generator winding temperature, respectively. Figure 2c illustrates the active power output as a function of wind speed of a fault-free turbine for a further reference. As Figure 2a demonstrates, many measured values of power output fall well inside the range of the normal power curve; so these measurements are defined as normal. However, when the wind speed is higher than 13 m/s, the turbine was operating for some periods of time with a power output reduced to around 1.5 MW and a generator winding temperature reduced to around 58 °C. These measurements indicate the wind turbine is operating in a fault condition. There are some discrete points deviating from the curves which are likely associated with different types of alarms. In our study, in order to investigate the periods of reduced power output and reduced generator winding temperature, only those data samples for which wind speed is higher than 13 m/s are considered. This would be beneficial as the general objective of the modelling process is to identify faults by comparing differences between the normal and the abnormal operational signals. Thus, for each of the 52 sensors used in this study, as

explained above, there are 5002 measurements remaining for which wind speed is higher than 13 m/s. The input data of the sensor array can be described as:

$$\Theta_1 = \left\{ \mathbf{X}^{(1)} \in \mathbf{R}^{52\times 5002}, x_{ij}^{(1)} \in \mathbf{R} \middle| i = 1, 2, \cdots, 52; j = 1, 2, \cdots, 5002 \right\} \quad (1)$$

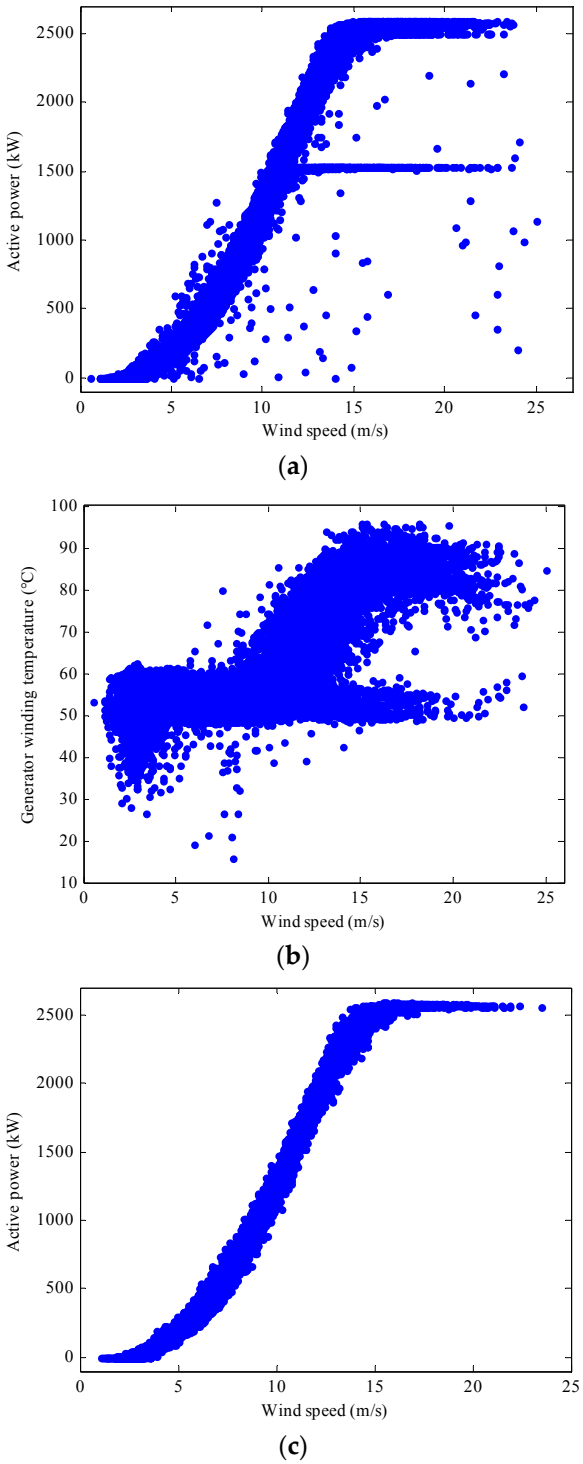


Figure 2. Scatter plots of the processed SCADA data: (**a**) Active power against wind speed; (**b**) Generator winding temperature against wind speed; (**c**) Active power output of a fault-free wind turbine.

3. PARAFAC Model

3.1. The Model

The notation and terminology to describe matrices and higher order arrays are adapted from [14]. Scalars are indicated by lower-case italics (e.g., x_{ijk}), vectors by bold lower-case characters (e.g., $\mathbf{y}$). Bold capitals ($\mathbf{X}$) are used for ordinary two-way data arrays (*i.e.*, data matrices) and underlined bold capitals ($\underline{\mathbf{X}}$) for three-way arrays. The letter I, J, K are reserved for indicating the dimensions of different modes. The ijkth element of $\underline{\mathbf{X}}$ is called x_{ijk}, where the indices can change in the following ranges: $i = 1, 2, \cdots, I$; $j = 1, 2, \cdots, J$; $k = 1, 2, \cdots, K$.

PARAFAC is a decomposition method, which conceptually can be compared to bilinear PCA, or rather it is one generalisation of bilinear PCA, while the Tucker3 decomposition is another generalization of PCA to higher orders [15]. The data are decomposed into triads or trilinear components; each component consists of one score vector and two loading vectors. It is common for three-way practice not to distinguish between scores and loadings as they are treated equally from a numerical perspective.

A PARAFAC model of a three-way array $\underline{\mathbf{X}}$ is given by three loading matrices, $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$. $\underline{\mathbf{X}}$ ($I \times J \times K$) can be written as follows:

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \tag{2}$$

Or in the form of a matrix [16]:

$$\mathbf{X} = \mathbf{A}\mathbf{T}^{(F \times FF)} (\mathbf{C} \otimes \mathbf{B})^{\mathrm{T}} + \mathbf{E} \tag{3}$$

where F is the number of factors which contribute to the signal; a_{if}, b_{jk}, c_{kf} are the elements of the loading matrices $\mathbf{A}(I \times F)$, $\mathbf{B}(J \times F)$, $\mathbf{C}(K \times F)$, respectively; e_{ijk} is the element of three-way residual data array of $\underline{\mathbf{E}}$ ($I \times J \times K$). The matrix $\mathbf{X}$ is $\underline{\mathbf{X}}$ rearranged to an $I \times JK$ matrix. The operator $\otimes$ is the so-called Kronecker product. The core array $\underline{\mathbf{T}}$ is a three-way array with zeros in all places except for the superdiagonal which contains ones, that is $t_{fgh} = 1$ for $f = g = h$, else $t_{fgh} = 0$. The superscript T stands for matrix transpose. The two-way matrix $\mathbf{T}^{(F \times FF)}$ is the $F \times FF$ metricized three-way core. The PARAFAC model may equivalently be stated in the form of a restricted Tucker3 model [17] as:

$$x_{ijk} = \sum_{f=1}^{L} \sum_{g=1}^{M} \sum_{h=1}^{N} a_{if} b_{jg} c_{kh} g_{fgh} + e_{ijk} \tag{4}$$

where $L = M = N = F$; the Tucker3 core $\underline{\mathbf{G}} = \underline{\mathbf{T}}$. If data $\underline{\mathbf{X}}$ is arranged in a three-way array (e.g., samples $\times$ times $\times$ sensors), the three-way PARAFAC model, as described in Equations (2) and (3), is shown graphically in Figure 3.

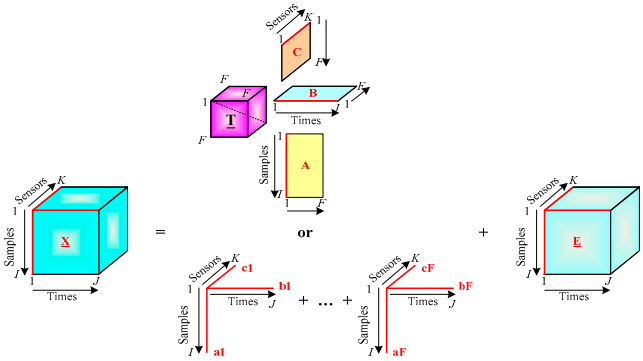


Figure 3. Graphical description of the PARAFAC model of $\underline{\mathbf{X}}$ (samples × times × sensors).

The trilinear model is found to minimise the sum of squares of the residuals (SSR) in the model:

$$SSR = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} e_{ijk}^2 \tag{5}$$

The alternating least squares (ALS) algorithms is used [16], which is based on the idea of reducing the optimisation problem to smaller sub-problems that are solved iteratively until the variation of the loss function or of the parameters is less than a predefined convergence criterion. An obvious advantage of the PARAFAC model is the uniqueness of the solution when there are not highly collinear components in the data [18].

3.2. The Core Consistency Diagnostic

The PARAFAC algorithm is very sensitive to F. If F is estimated to be too low, there is no physical meaning. If F is too high, the noise will be increasingly modelled and the true factors will be modelled by more correlated components. The core consistency diagnostic is used as an indicator of an appropriate PARAFAC model. It may quantify the similarity between $\underline{\mathbf{G}} = \underline{\mathbf{T}}$:

$$\delta = 1 - \frac{\sum_{d=1}^{F} \sum_{e=1}^{F} \sum_{f=1}^{F} (g_{def} - t_{def})^2}{F} \tag{6}$$

The core consistency δ is always less than or equal to 100% and may also be negative. The value of δ above 80% implies an appropriate model achieved, whereas a core consistency in the neighbourhood of 50% means an inaccurate model. δ close to zero or even negative implies an invalid model because the space covered by the loading matrices is then not primarily describing trilinear variation [16].

4. K-means Clustering Method

In order to verify if the extracted features from the PARAFAC model are good for system identification, K-means clustering method is used. Essentially, K-means clustering automatically divides a data set into K groups [19]. It proceeds by selecting k initial cluster centres and then iteratively refining them as follows.

Let $x = \{x_i\}$, $i = 1, \ldots, n$ be the set of n dimensional observations to be clustered into a set of K clusters $c = \{c_k\}$, $k = 1, \ldots, K$. K-means algorithm finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimised [20]. Let u_k be the mean of cluster c_k. The squared error between u_k and the points in cluster c_k is defined as:

$$J(c_k) = \sum_{x_i \in c_k} ||x_i - u_k||^2 \tag{7}$$

The goal of K-means is to minimise the sum of the squared error over all K clusters:

$$J(\mathrm{c}) = \sum_{k=1}^{K} \sum_{x_i \in c_k} ||\, x_i - u_k \,||^2 \tag{8}$$

Minimising this objective function is known to be an non-deterministic polynomial (NP)-hard problem (even for $K = 2$) [21]. Thus K-means, which is a greedy algorithm, can only converge to a local minimum. One way to overcome the local minima is to run the K-means algorithm, for a given K, with multiple different initial partitions and choose the partition with the smallest squared error. K-means is used with the Euclidean metric for computing the distance between points and cluster centres. Therefore, K-means tends to find spherical or ball-shaped clusters in data.

5. Results and Discussion

5.1. Data Pre-Processing

In order to build the PARAFAC model, the data array Θ_1 is first arranged into a three-way array $\underline{\mathbf{X}}$ ($I \times J \times K$) of three dimensions as samples $\times$ times $\times$ sensors. In order to examine whether each measurement value is indication of normal or abnormal operation, the first dimension is arranged with the 5002 measurement samples and the second dimension is only one associated with time. The third dimension is the 52 sensors, thus $\underline{\mathbf{X}}$ (5002 $\times$ 1 $\times$ 52) is obtained:

$$\Theta_2 = \left\{ \underline{\mathbf{X}} \in \mathbf{R}^{5002 \times 1 \times 52},\, x_{ijk} \in \mathbf{R} \,\middle|\, i = 1, 2, \cdots, 5002; j = 1; k = 1, 2, \cdots, 52 \right\} \tag{9}$$

One element x_{ijk} in $\underline{\mathbf{X}}$ corresponds to the measured sensor signal value i at time instant j from sensor k. For each sensor, every sample represents a value measured at a particular time instant. The measurement data can be considered as linearly independent of each other because many processes associated with wind turbines are non-linear and measurements are made at 10 min intervals. The sensors are used to monitor different phenomena including electrical, mechanical, thermal, chemical and meteorological phenomena, which means that the components will not be highly collinear in dimension three. Thus the factors of the PARAFAC model can be uniquely determined [18].

In order to improve the accuracy for fault detection, mean-centring is performed, which aims to remove constant terms in the dataset in order to increase the difference between the samples. Mean-centring across the first dimension (*i.e.*, the sample dimension) can be done by metricising the array to an $I \times JK$ matrix, and then centring it in an ordinary two-way analysis [22]:

$$x_{ijk}^{\text{centered}} = x_{ijk} - \sum_{i=1}^{I} x_{ijk}/I \tag{10}$$

5.2. Determining the Factors

To determine the factors, PARAFAC models with an increasing number of factors from 1 to 5 are built based on the mean-centred data. The value of the core consistency is 100%, 98%, 873.%, 44.2% and 31.7%, respectively. It is typically decreasing more or less monotonically with the number of factors, because the influence of noise and other non-trilinear variation increase with number of factors. It is found that the number of factors, which better describe the data, should be $F = 3$. The three-factor PARAFAC model has a core consistency of 87.3%. When the factor is over 3, it will lead to a sharp decrease in the degree of core consistency.

The PARAFAC model with $F = 3$ can be obtained by using ALS algorithm, as described in Equation (5). Consequently, the three loading matrices are therefore $\mathbf{A}$ (5002 $\times$ 3), $\mathbf{B}$ (1 $\times$ 3) and $\mathbf{C}$ (52 $\times$ 3), corresponding to the sample mode, the response mode and the sensor mode, respectively. The number of columns of the three matrices are all 3. In the subsequent sections, the different modes

will be noted as follows: sample mode (**A**): **a**1, **a**2, **a**3; response mode (**B**): **b**1, **b**2, **b**3; sensor mode (**C**): **c**1, **c**2, **c**3. The sample mode and sensor mode will be illustrated with scatter loadings plots.

5.3. Fault Detection

Fault detection aims to identify data patterns which do not conform to the principle of expectation. The loading plot (Figure 4) showing correlations between the first two factors (**a**1 *versus* **a**2) for the sample mode (the loading matrix **A**) offers a clear way for visualising all the samples. Because mean-centring across the first dimension was done as described in Equations (10), most of measurement points are accumulated in the vicinity of the coordinate origin (0,0). Compared with Figure 2, it can be seen that these measurement points represent normal operation of the wind turbine. Some of them are clustered in the left plot far away from the origin indicating fault conditions with a reduced active power output, while the points scattered in the lower part of the plot representing alarm conditions.

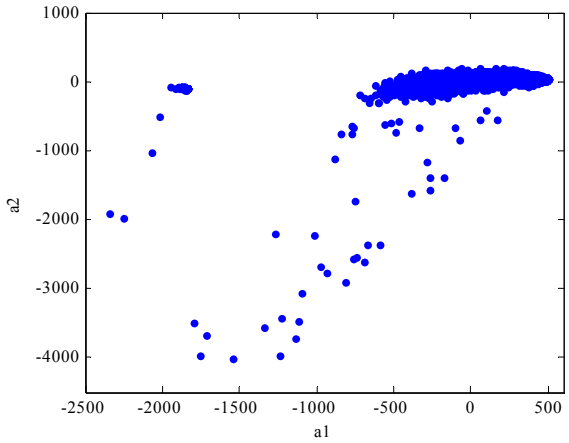


Figure 4. Loadings plot of the PARAFAC model for the sample mode (**a**1 *versus* **a**2), **a**1 and **a**2 are the first two columns of the loading matrix **A.**

In order to verify if the extracted features by the PARAFAC model are good ones for fault detection and to find out which group each measurement point belongs to, the loading matrix **A** is now analysed using *K*-means clustering method. Figure 5 shows that three types of measurement data are clearly distinguished from each other. The alarm samples distributes more dispersedly, which indicates the great variability of their measurement values.

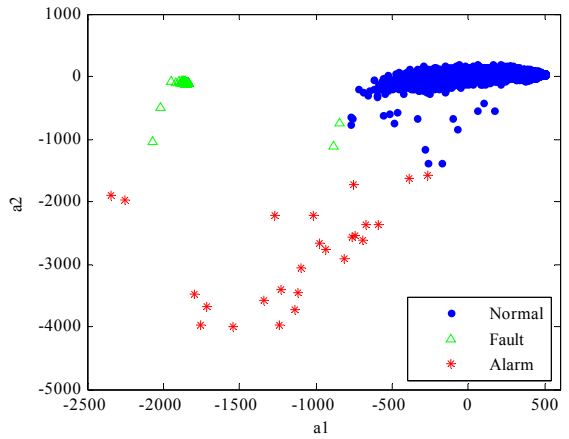


Figure 5. *K*-means clustering result of the loading matrix **A**, **a**1 and **a**2 are the first two columns of **A**.

Each sample represents a value measured at a specific time, so we can know which group it belongs to. Thus the active power and the generator winding temperature can be plotted against measurement time and wind speed, as shown in Figures 6 and 7 respectively. In these figures, the blue dots represent normal measurement points while the green triangles represent fault points, and the red stars represent alarm points. These figures indicate a clear distinction between the three operational patterns of the turbine. Having checked the alarm logs, all the measurement points for which the active power is lower than 1.5 MW while the wind speed is larger than 13 m/s are marked with red stars in Figure 6. These figures also show several significant features, as summarised below:

1. The active power normally increases with increasing wind speed until it reaches a stable value of approximately 2.5 MW. The generator winding temperature also increases with the rising wind speed.
2. During the fault operation period, the active power and the generator winding temperature are reduced to around 1.5 MW and 50–58 °C, respectively, despite the increase in wind speeds. This time period occurs from sample 3711 to sample 4162 for a duration of 4510 min. It was found from the alarm logs that a low gearbox oil sump temperature appears to have been the cause of the problem.
3. The discrete points marked as red stars in Figure 6 indicate short excursions of the active power outside the normal range. It was found from the alarm logs that the problems were mostly associated with the wind speed which was close to the cut-out speed and with high or low gearbox oil sump temperatures.

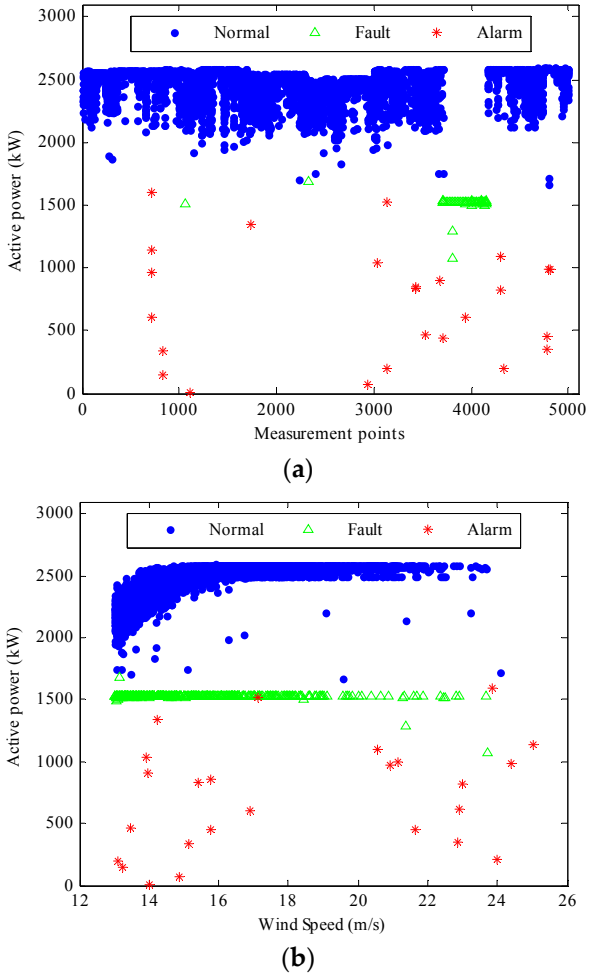


Figure 6. Active power output for fault detection: (**a**) Active power output against measurement time; (**b**) Active power output against wind speed.

For example, at samples 4783 and 4784, the wind speed was 22.9 and 21.7 m/s, respectively, which might be too strong for the generator and can thereby damage the turbines. Thus, the brake device has to be put in use to stop the wind turbine from running at high wind speeds. In sample 3434, the gearbox oil sump temperature reached a high value of up to 73.88 °C. In sample 2944, it was reduced to 47.52 °C. In order to ensure the safe operation of the turbine, the system automatically limits the power operation when the gearbox oil temperature is too high or too low. The discrete generator winding temperature points, as shown in Figure 7, are also related to these wind speed and gear oil sump temperature events.

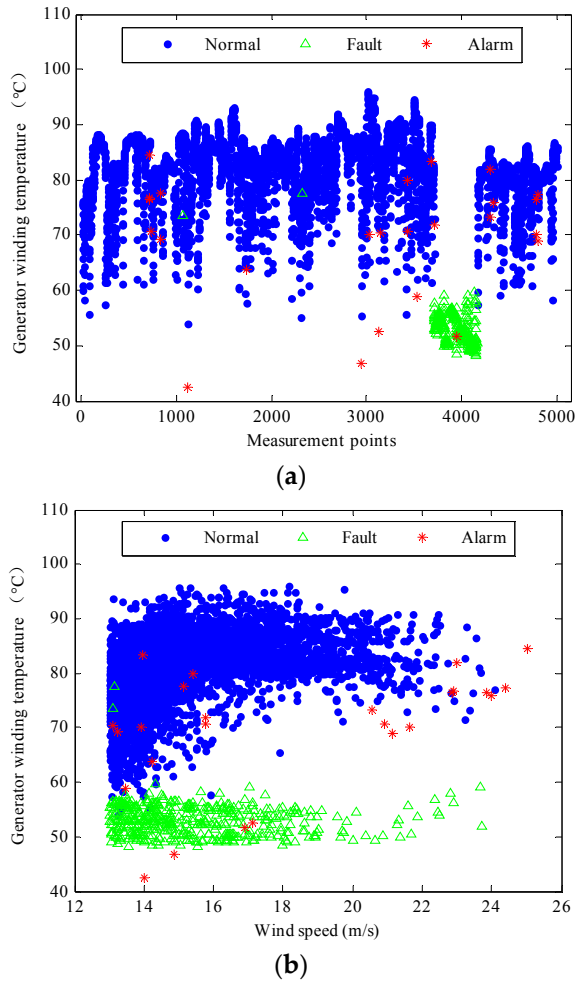


Figure 7. Generator winding temperature for fault detection: (**a**) Generator winding temperature against measurement time; (**b**) Generator winding temperature against wind speed.

5.4. Sensor Selection

With regards to the sensor mode (the loading matrix **C** which is relative to the sensor contributions), the loading plot **c**1 *versus* **c**2 and an enlarged view of the central cluster are shown in Figure 8. It can be seen that 52 sensors can be divided into eight groups, which are summarized in Table 1. Most of the sensors (Group VIII) are concentrated in the vicinity of the coordinate origin (0,0). Sensor 1 (Group I), sensors 26–29 (Group II), sensors 9–11 (Group III), sensors 12–13 (Group IV), sensors 14 (Group V), sensors 2–8 (Group VI) and sensors 33, 34, 48 (Group VII) belong to the different groups, respectively.

Table 1. The eight groups of the sensors.

Group	The Number of Sensors	Parameter Description
I	1	Nacelle position
II	26–29	Generator speeds and the rotor speed
III	9–11	Mains currents
IV	12–13	Apparent power and active power
V	14	Reactive power
VI	2–8	Blade positions
VII	33,34,48	Generator temperatures
VIII	others	Other parameters

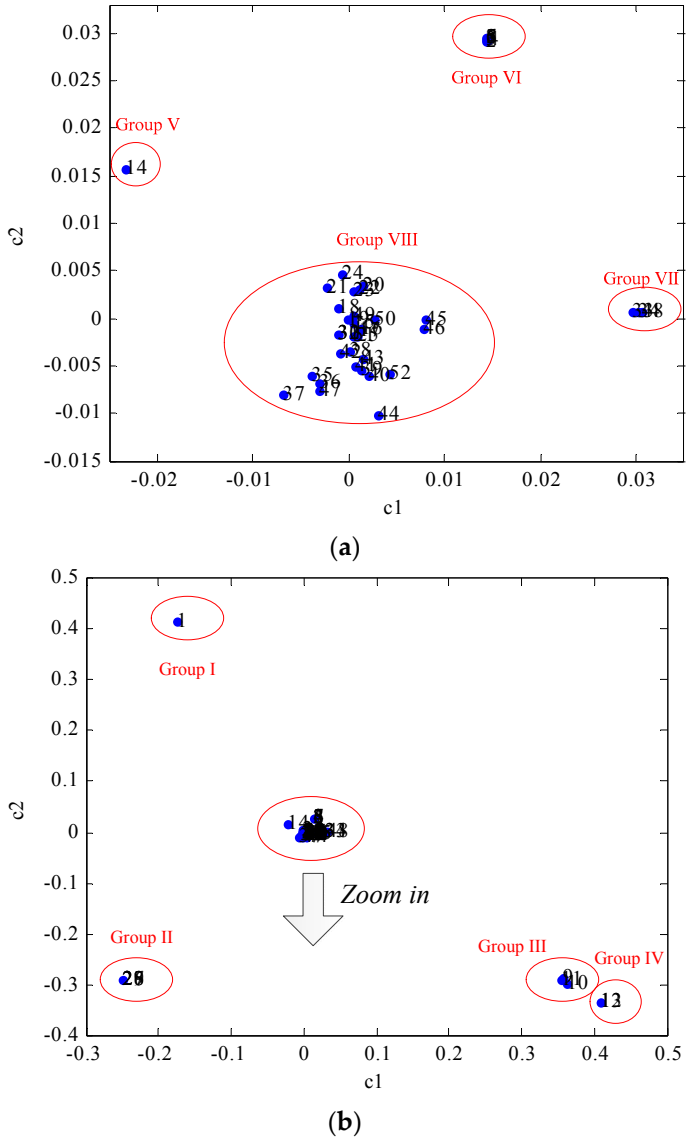


Figure 8. Loadings plot of the PARAFAC model for the sensor mode (**c**1 *versus*. **c**2): (**a**) Loadings plot of the PARAFAC model for sensor mode; (**b**) An enlarged view of the central cluster of (**a**) , **c**1 and **c**2 are the first two columns of the loading matrix **C**.

The readings of the grouped sensors generally behave similar to each other and are highly correlated. Group I-VII lie in the areas far from the origin, which indicates that the response values of these sensors change notably during a fault condition. On the contrary, the signals of Group VIII changed a little, which means that the fault condition does not cause these signals to change greatly during the fault. For example, during the fault operation of the wind turbine, the active power (sensor 13) and generator winding temperature (sensor 33) decrease greatly, as shown in Figures 6a and 7a, respectively, while the gearbox oil sump temperature (sensor 47) remained relatively constant, as shown in Figure 9.

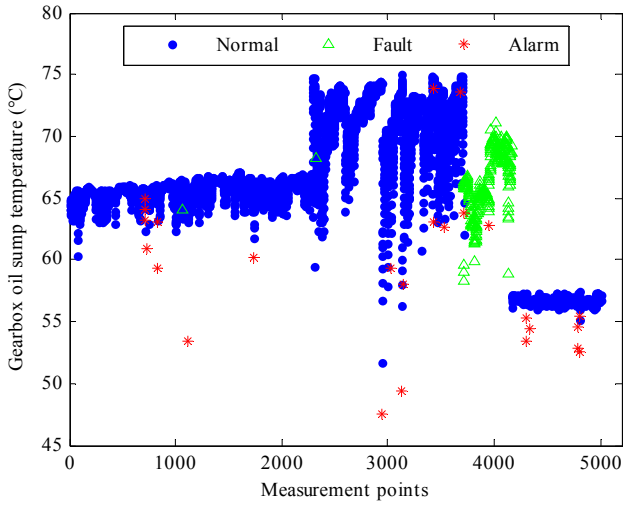
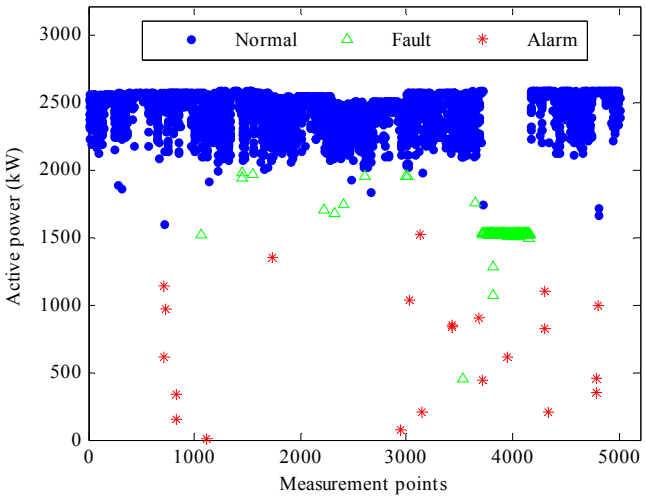


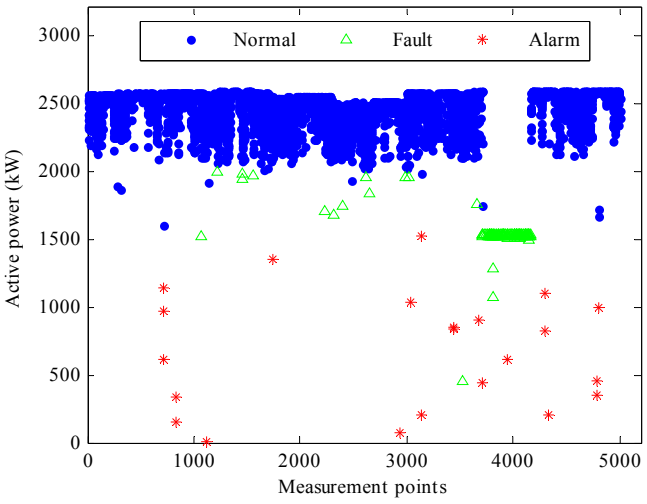
Figure 9. Gearbox oil sump temperature for fault detection.

We can randomly choose one sensor from each respective group. Therefore under this wind turbine operation condition, the size of the sensor array can be reduced to eight by selecting sensors as shown in Table 1. If sensors 1 (nacelle position), 2 (pitch position 1), 9 (mains current phase A), 12 (apparent power), 14 (reactive power), 26 (generator speed), 33 (temperature generator 1), 52 (temperature cooling water) are selected, the responses of these eight sensors can be used to characterize the whole sensor array response pattern under this operation condition. With the method described in Section 5.3, the PARAFAC model of this optimised array can be obtained, and the active power output against measurement time for fault detection is plotted in Figure 10a.



(**a**)

Figure 10. *Cont.*



(**b**)

Figure 10. Active power output produced by PARAFAC model from samples using an optimised sensor array: (**a**) Using sensors 1, 2, 9, 12, 14, 26, 33, 52; (**b**) Using sensors 1, 8, 11, 13, 14, 19, 29, 48.

6. Conclusions

The use of PARAFAC is studied for condition monitoring of wind turbines, evaluated with SCADA data from an operational wind farm. In order to build a suitable model, the SCADA data are selected and pre-processed in order to obtain an appropriate three-way dataset. The results from the PARAFAC model show its effectiveness in providing easily interpretable plots revealing the turbine conditions. Combined with the *K*-means clustering method, three kinds of wind turbine operation conditions are identified, *i.e.*, normal, fault and alarm conditions. In the meanwhile, the contribution results of the sensors are also utilized to optimise the sensor array and to reduce data redundancies between sensors. The sensors are classified into eight groups. By selecting one representative senor randomly from each group, the eight sensors are able to characterize the whole sensor array response pattern. It can be found from the loading plot for the sensor mode that there were remarkable changes in the powers, main currents, and generator winding temperatures when the turbine operated under fault conditions. However, further work remains to determine the validity of the approach, including testing the measurement data from more wind turbines. The measurements when the wind speed is relatively low (less than 13 m/s in this study) must also be considered so that results can be more robust and convincing, allowing decisions to be made for optimal maintenance scheduling.

Acknowledgments: This research work is supported by Chinese Scholarship Council and the support of Engineering Department at Lancaster University. The permission of use SCADA data from Wind Prospect Ltd. is also gratefully acknowledged.

Author Contributions: Wenna Zhang carried out most of the work presented here, Xiandong Ma revised the contents and reviewed the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Herbert, G.M.J.; Iniyan, S.; Sreevalsan, E.; Rajapandian, S. A review of wind energy technologies. *Renew. Sust. Energ. Rev.* **2007**, *11*, 1117–1145. [CrossRef]
2. Hameed, Z.; Hong, Y.S.; Cho, Y.M.; Ahn, S.H.; Song, C.K. Condition monitoring and fault detection of wind turbines and related algorithms: A review. *Renew. Sust. Energ. Rev.* **2009**, *13*, 1–39. [CrossRef]
3. Alix, K.; Lariagon, C.; Delourme, R.; Manzanares-Dauleux, M.J. A model-based approach to wind turbine condition monitoring using SCADA data. *Plant Breed.* **2009**, *38*, 536–548.

4. Marvuglia, A.; Messineo, A. Monitoring of wind farms' power curves using machine learning techniques. *Appl. Energ.* **2012**, *98*, 574–583. [CrossRef]
5. Cross, P.; Ma, X. Model-based and fuzzy logic approaches to condition monitoring of operational wind turbines. *Int. J. Autom. Comput.* **2015**, *12*, 25–34. [CrossRef]
6. Ma, X. Novel Early Warning Fault Detection for Wind-turbine-based DG Systems. In Proceedings of the 2nd IEEE PES International Conference and Exhibition on Innovative Smart Grid Technologies (ISGT Europe), Manchester, UK, 5–7 December 2011; pp. 1–6.
7. Kroonenberg, P.M. *Three-Mode Principal Component Analysis: Theory and Applications*; DSWO Press: Leiden, The Netherlands, 1983.
8. Bartholomew, D.J. Analysis and Interpretation of Multivariate Data. In *International Encyclopedia of Education (Third Edition)*; Elsevier: Oxford, UK, 2010; pp. 12–17.
9. Harshman, R.A. In Foundations of the PARAFAC procedure: Model and conditions for an "explanatory" multi-mode factor analysis. *UCLA Work. Pap. Phon.* **1970**, *16*, 1–84.
10. Padilla, M.; Montoliu, I.; Pardo, A.; Perera, A.; Marco, S. Feature extraction on three way enose signals. *Sens. Actuators B Chem.* **2006**, *116*, 145–150. [CrossRef]
11. Phan, A.H.; Cichocki, A. PARAFAC algorithms for large-scale problems. *Neurocomputing* **2011**, *74*, 1970–1984. [CrossRef]
12. Baunsgaard, D.; Andersson, C.A.; Arndal, A.; Munck, L. Multi-way chemometrics for mathematical separation of fluorescent colorants and colour precursors from spectrofluorimetry of beet sugar and beet sugar thick juice as validated by HPLC analysis. *Food Chem.* **2000**, *70*, 113–121. [CrossRef]
13. Andersen, C.M.; Bro, R. Practical aspects of PARAFAC modeling of fluorescence excitation-emission data. *J. Chemom.* **2003**, *17*, 200–215. [CrossRef]
14. Bro, R. PARAFAC. Tutorial and applications. *Chemom. Intell. Lab. Syst.* **1997**, *38*, 149–171. [CrossRef]
15. Harshman, R.A.; Berenbaum, S.A. Appendix A—Basic Concepts Underlying the PARAFAC-CANDECOMP Three-Way Factor Analysis Model and its Application to Longitudinal Data 1,2. In *Present and Past in Middle Life*; Clausen, A.A., Haan, N., Honzik, M.P., Mussen, P.H., Eds.; Academic Press: Cambridge, MA, USA, 1981; pp. 435–459.
16. Bro, R.; Kiers, H.A.L. A new efficient method for determining the number of components in PARAFAC models. *J. Chemom.* **2003**, *17*, 274–286. [CrossRef]
17. Bro, R. Multi-way Analysis in the Food Industry. Ph.D. Thesis, Royal Veterinary and Agricultural University, Frederiksberg, Denmark, 1998.
18. Kruskal, J.B. Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra Appl.* **1977**, *18*, 95–138. [CrossRef]
19. Honda, K.; Notsu, A.; Ichihashi, H. Fuzzy PCA-Guided Robust k-Means Clustering. *IEEE Trans. Fuzzy Syst.* **2010**, *18*, 67–79. [CrossRef]
20. Jain, A.K. Data Clustering: 50 Years Beyond K-means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666. [CrossRef]
21. Drineas, P.; Frieze, A.; Kannan, R.; Vempala, S.; Vinay, V. Clustering Large Graphs via the Singular Value Decomposition. *Mach. Learn.* **2004**, *56*, 9–33. [CrossRef]
22. Rasmus, B.; Smilde, A.K. Centering and scaling in component models. *J. Chemom.* **2003**, *17*, 16–33.