Proceedings of Global Power and Propulsion Society
ISSN-Nr: 2504-4400
GPPS Chania24
September 4-6, 2024
www.gpps.global

GPPS-TC-2024-0101

NORMAL BEHAVIOUR MODELING OF HAWT FLEETS USING SCADA-BASED FEATURE ENGINEERING

Valerio F. Barnabei
University of Rome La Sapienza
valerio.barnabei@uniroma1.it
Rome, Italy

Filippo De Girolamo
University of Rome La Sapienza
filippo.degirolamo@uniroma1.it
Rome, Italy

Tullio C. M. Ancora
University of Rome La Sapienza
tulliocarlo.ancora@uniroma1.it
Rome, Italy

Lorenzo Tieghi
University of Rome La Sapienza
lorenzo.tieghi@uniroma1.it
Rome, Italy

Giovanni Delibra
University of Rome La Sapienza
giovanni.delibra@uniroma1.it
Rome, Italy

Alessandro Corsini
University of Rome La Sapienza
alessandro.corsini@uniroma1.it
Rome, Italy

ABSTRACT

Wind turbine O&M presents a challenge for large wind turbine fleets, where many turbines are scattered across different wind farms, exposed to non-homogeneous environmental conditions and varied operational histories. In this scenario, operators can leverage big data to address this challenge, with the objective of deriving O&M strategies common to subsets or communities of turbines, either to monitor their behavior collectively or to schedule custom maintenance.

In this paper, we introduce an unsupervised data-driven framework to isolate communities of wind turbines based on their actual operational histories, as a precursor to asset-specific O&M strategies. The approach is based on a novel Meta Predictive Power Score (MPPS) to derive a behavior model for horizontal axis wind turbine fleets. The proposed framework relies on feature derivation, feature combination, and community detection algorithms.

The feature derivation component involves the use of a multivariate feature selection algorithm based on the Combined Predictive Power Score (CPPS), in which a regression task quantifies the information content of combinations of variables for the prediction of one or more target parameters. A feature combination procedure defines the core of MPPS, where in a combination of CPPS scores and selected features for different regression tasks on diverse versions of the same turbine SCADA data results in a similarity matrix for the entire fleet.

A community detection algorithm, based on Complex Network Analysis, is then used to identify groups of wind turbines within a fleet that exhibit similar behavior. The dataset used comprises 67 wind turbines, including diverse OEMs. The proposed algorithm employs a decision tree regressor considering lagged time series to compute the MPPS. The results demonstrate the flexibility of the multi-input multi-output formulation. The target of the regression tasks required to build the scores is a hypersurface that can describe the operational state of a wind turbine, defined by active power, blade pitch angle, and rotational speed.

INTRODUCTION

Wind power represents, in all possible technology exploitation i.e. on-shore and more recently off-shore and high-altitude concepts, the backbone of energy transitioning policies in western countries as well as in the far east. In view of the instrumental role of wind power in the decarbonization process, European offshore wind capacity will grow from its current level of 12 GW to at least 60 GW by 2030 and to 300 GW by 2050 An (2020), and similar trends are consolidating worldwide. The increase in wind farm producibility is pursued by installing larger turbines, in more dense farm layouts, in more severe operating environments (higher height, off-shore in deep sea). From an O&M perspective, all these factors contribute to producing new and unforeseen (based on track records) loading patterns affecting the aero-hydro-servo-elastic behavior of wind turbine rotors, resulting in even more severe operations of the drive-trains Kong et al. (2021). Diagnosis and prognosis of incipient faults in drive-train components can effectively reduce the economic loss caused by downtime in power generation and unplanned maintenance, repairs, and replacements. O&M challenges in wind energy advocate condition monitoring strategies capable of early detection and isolation of incipient faults by using diagnostic routines to enable so-called condition-based maintenance (CBM) approaches. The task of operation monitoring is customarily based on a network of sensors that are part of Supervisory Control and Data Acquisition (SCADA) systems meant to collect, store, and display all relevant parameters of individual wind turbines and farm components. In the wind technologies arena, SCADA data are usually collected on a 10-minute interval basis to provide statistical measures (mean, standard deviation, maximum value, and minimum value) of signals, to derive energy output measures, turbine availability, and error logs Wen and Xie (2021). As a consequence, SCADA systems record vast quantities of raw data, which may suffer from a certain degree of unreliability with inaccurate and incomplete entries, missing values, or noisy sensor measurements Chen et al. (2021). All those abnormal entries may directly impact the development of data-driven strategies dedicated to power output control and prediction Qiao et al. (2021), or anomaly detection Jin et al. (2020). Due to the fact that wind energy assets consist of numerous components with possibly different operating histories, component-specific behavior definition may only rely on data-driven tools designed to integrate feature engineering approaches with daily operating schedules Ibrahim et al. (2011). The ability to predict, with good accuracy, wind farm operations or to anticipate faults to drive CBM, can be based upon the discovery of normal behavior models (NBM). In large WT fleets, this insight into NBM may enable the exploitation of high-frequency SCADA datasets to nowcast potential faults and forecast producibility to manage responsive wind farm operations. Learning algorithms are nowadays considered the most effective way to spot operational inefficiencies, improve producibility, and forecast power outputs for grid balancing purposes. However, there are open issues directly linked to the nature of the data and to the specific task. First, large SCADA datasets lead to challenges in training Machine Learning models due to inevitable redundancies, high computational requirements, and lack of homogeneity in the variable dictionaries, mostly in cases of fleets with diverse turbine Original Equipment Manufacturers (OEMs). Second, it is still questionable whether wind turbine normal behavior models can be derived directly from the SCADA features to establish asset management strategies for wind turbine fleets rather than for single units without the adoption of complex transfer learning methods Jin et al. (2022) or relying on multiple models Meyer (2021). A common strategy is to derive single-target NBMs, with a trade-off on the possible selection of monitored variables and the descriptive power of the NBM, while a multi-target strategy can result in a more reliable description of the NBM Meyer (2021). Finally, the definition itself of normal behavior usually requires a deep understanding of the specific wind turbine to isolate an operating region or period in time to derive such NBM; the results, however, can be deeply affected by the different operational history of the wind turbine, which is not always brand new when purchased. To this end, an unsupervised approach like clustering Izquierdo et al. (2019) can be adopted. However, this can pose additional constraints on the methodology, given the homogeneity requirement in datasets to allow those methods to work. Alternatively, a community detection algorithm to maximize modularity Brandes et al. (2007) in a complex system (like a large non-homogeneous wind turbine fleet) appears promising. A community can be derived within a graph representation of a network, by the emergence of some correlation among the nodes of the graph (i.e. the wind turbines in a fleet), with such correlation interpreted as the edge of the graph. To get this correlation, it is possible to perform a feature derivation task to condense the information available in SCADA data into new quantities, usually scalars (i.e. abandoning the time series domain) that could relate a couple of turbines. A graph representation of such a correlation matrix is then explored with community detection algorithms, also to further explore these new relationships dynamically. In this paper, we propose a novel unsupervised framework to discover communities with comparable behavior of wind turbines, referred to as community NBM, based on SCADA datasets semi-supervised exploration. In detail, the NBM framework combines the derivation of features from the original SCADA data to create wind turbine model agnostic metrics, a wind turbine behavior model based on the aforementioned derived metrics, and a community detection algorithm rooted in the field of Complex Network Analysis. Concerning the feature derivation step, we advocate the use of Combined Predictive Power Score (CPPS), first introduced in Miele et al. (2022) as a metric to evaluate feature importance in multivariate problems, that outperforms other standard approaches in view of its multivariate nature and that was recently used in fault anticipation on WT drivetrain components Bonacina et al. (2022). An early attempt to adopt CPPS to derive a NBM on wind turbine fleets can be found in Barnabei et al. (2023), where the scores from a regression with only active power as the target still defined communities of similar behavior.

The proposed framework integrates the following stages: i. a feature derivation or creation task using the CPPS, to derive a new metric called Meta Predictive Power Score (MPPS); ii. an innovative method to derive a multi-target Normal Behavior Model (NBM) for wind turbines based on the CPPS features and metrics; iii. a community detection algorithm, based on Complex Network Analysis, to unveil groups of wind turbines (within a fleet) featuring similar behavior to define common O&M strategies and distinguish them based on the actual wind turbine fleet operational history.

The proposed approach can be intended as a support and enhancement of O&M strategies with multiple advantages: i. an operator could schedule custom maintenance for turbine subsets or monitor custom parameters for turbine subsets, even for turbines from the same manufacturer; ii. the unsupervised nature of the method allows the characteristic dynamics that differentiate the operational history and environmental conditions of the turbines to automatically emerge from the data; iii. the proposed method is set up to separate the different contributions of environmental conditions and turbine-specific

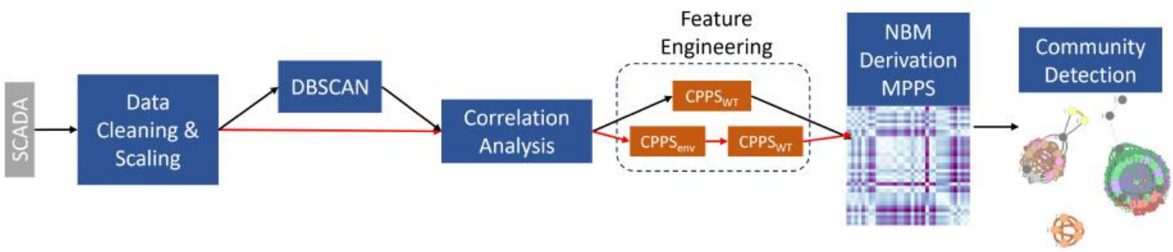


Figure 1 Proposed methodology data flow

operating conditions; iv. the core of the method is expected to be agnostic to the common data management issues in SCADA-based methods related to non-homogeneous SCADA datasets in terms of monitored variables.

The case study includes SCADA data collected from a fleet of 67 turbines of the same model and belonging to five wind farms located in the Central-Southern Apennines. The data, recorded every 10 minutes throughout a year of operation, comprise 36 monitored variables, including temperatures of gearbox, generator, and hydraulic systems, along with active power, current, and wind conditions. The outcomes of the suggested approach are examined in relation to the active power target variable.

METHODOLOGY

The methodology is designed to identify groups of wind turbines with similar behavior from SCADA data, regardless of their geographical positions. The goal is to use a novel feature derivation algorithm, which is multi-variate by definition, to determine new metrics concerning the behavior of wind turbines. Figure 1 outlines the sequence of processing steps from data collection to community detection. The proposed methodology includes four main components: data pre-processing, feature derivation or creation using CPPS, derivation of a wind turbine's normal behavior model using MPPS, and finally, a community detection algorithm.

In the first step, large outliers and missing values are removed from the single turbine SCADA dataset. Next, a density-based clustering algorithm runs iteratively on the target hypersurface to further isolate the characteristic and regular behavior of the turbine. This results in two distinct versions of the dataset: a raw dataset where no clustering is applied (red arrows), and a filtered dataset (black arrows). These two datasets undergo pre-processing via time series clustering to assess the correlation among variables.

Next, the CPPS algorithm defines the best predictors and a score for each version of the dataset in two separate steps: one using only environmental variables for the regression of the hypersurface, and the other using all the remaining wind turbine-related variables. This score measures how well a selected multivariate regressor outperforms a naive predictive model in accurately predicting a multi-dimensional output. The similarity matrix is then constructed, incorporating the values of the new metric, the Meta Predictive Power Score (MPPS), which combines the different CPPS values and selected features for each turbine.

A particular focus is given to evaluating how the predictive capability of the regressor is affected by the presence or absence of non-characteristic operating conditions. This is done by comparing the CPPS values obtained considering only the characteristic behavior of the turbine with those obtained by including points of anomalous operation. The final step features a community detection algorithm, based on Complex Network Analysis, to divide the turbines into communities characterized by similar behavior.

Data Preparation

Preparation of SCADA data and management of its anomalies represent the first and fundamental step for all machine learning and data analytics methods in this field. Anomalies removal, and thus the identification of the machine's characteristic curve, is addressed in the literature in various ways. Specifically, some methods are preceded by data pre-processing, while in others, curves are directly filtered by applying anomaly detection techniques to data in their original form Morrison et al. (2022). Approaches not preceded by pre-processing include Long et al. (2019), which uses image-based techniques, and Wang et al. (2014), which employs Gaussian Mixture Models. In this work, however, an initial statistical filter is applied, followed by DBSCAN, as also done in Zhao et al. (2017).

For each signal, a 6-hour sliding window is defined and samples outside the interquartile range are removed. Subsequently, a density-based clustering Ester et al. (1996) is applied to the three curves defined by the three target variables plotted against the wind speed velocity, excluding samples which are not part of the main curve's clusters. This leads to the creation of a filtered dataset composed only of samples which strictly represent the wind turbine's operational characteristic curves. Excluded SCADA entries from the clustering procedure mainly represent power curtailment requests from the grid, transient dynamics at cut-in, and in general, every entry in the wind turbine SCADA where the measured power or wind turbine regime was not consistent with the wind conditions. On the other hand, a raw counterpart of the dataset, which also includes such non-characteristic operational conditions, is preserved and used to derive the MPPS.

A quantity representing the ambient turbulence intensity (TI) is added to both dataset, defined as:

$$TI = \frac{v_{sd}}{v} \tag{1}$$

where v_{sd} and v are, respectively, the wind speed standard deviation and the wind speed average in the last 10 minutes.

In order to avoid highly correlated variables being selected as important predictors for each other both datasets variables are scaled using the Robust Scaler Pedregosa et al. (2011) and then grouped in homogeneous groups, as proposed by Bonacina et al. (2022).

Feature derivation

CPPS is introduced in Miele et al. (2022) as an augmented Predictive Power Score (PPS) method Wetschoreck et al. (2020), which is a metric to evaluate feature importance in multivariate problems and outperforms other standard approaches due to its multivariate nature. PPS, a sequential selection algorithm, is originally proposed as an asymmetric, data-type-agnostic score ranging from 0 to 1 to detect linear or nonlinear relationships between two variables. It is calculated using one variable at a time to predict another one (within the full set of features). CPPS extends the original formulation of PPS by considering combinations of candidate input features $F_g = \{f_i\} \subseteq F$ for the prediction of multiple target features $T = \{f_T\}$, where F is the set containing all input features, f_i is a candidate input, and f_T is a target. Therefore, CPPS is designed to detect linear or non-linear relationships between different groups of variables by significantly speeding up PPS computational time. CPPS compares regressor performance against a naive regressor, for a candidate group of variables F_g and a target group T:

$$CPPS(F_g, T) = 1 - \frac{MAE^{model}_{F_g,T}}{MAE^{naive}_{T}} \tag{2}$$

where $MAE^{model}_{F_g,T}$ is the Mean Absolute Error (MAE) of the regression model while MAE^{naive}_{T} is that of the median of each target.

In this study, the regressor is a Decision Tree-based regression model, and three variables define T: active power, revolutions per minute, and blade pitch angle. The target of the multivariate regressor is thus a hyper-surface in the space of the target variables, as shown in Figure 2 for the raw dataset. Figure 2 shows the normalized values of RPM and blade pitch angle, and a dimensionless active power, scaled with respect to the nominal power of the wind turbine.

In order to consider the output of the time series clustering algorithm, for each feature f_i added to a new combination F_g, all variables belonging to the same homogeneous groups are appended to the combination.

The CPPS algorithm is performed on both datasets. The raw dataset results in two separate contributions: $CPPS_{env}$, evaluated only using environmental variables (i.e., wind direction, wind speed, ambient temperature, ambient turbulent intensity) for the regression of the hypersurface, and $CPPS^{raw}_{WT}$, evaluated only using quantities representing drive-train and generator status (e.g., drive end bearing temperatures, generator bearing temperature, and hydraulic pressure, among others) for the regression of the hypersurface. In the attempt to define a turbine-specific behavior, $CPPS_{env}$ only accounts for the score itself, as all the environmental features are used in the regression. $CPPS^{raw}_{WT}$, on the other hand, accounts for both the score and the list of selected predictors for the regression.

From the filtered dataset, which represents the turbine's operational characteristics, scores are calculated based solely on machine variables ($CPPS^{filt}_{WT}$).

Normal Behaviour Model

A wind turbine behavior model can be derived from the CPPS feature selection procedure, building on findings from Bonacina et al. (2022), where CPPS resulted in a condensed quantity to evaluate for turbine stability, and identify incipient faults. This implies that CPPS scores, combined with the most important predictors, can form a foundation to represent turbine status, ensuring stability among turbines in a wind farm. It is therefore possible to define the new Meta Predictive Power Score (MPPS) metric, to define a normal behaviour model for a fleet of $i, j = 1, \ldots N$ WTs. MPPS is a score defined for a couple of turbines i, j and represents, the contribution of multiple factors in the attempt to perform regression of the target hypersurface of Figure 2, and it's filtered counterpart. The set of scores emerging by comparing each WTs couple in the fleet allows the definition of a MPPS matrix, that can be interpreted as a similarity matrix. An element $MPPS_{ij}$ of the similarity matrix is defined as:

$$MPPS_{ij} = [E_{ij} + (A_{ij} \cdot C_{ij})]^n \tag{3}$$

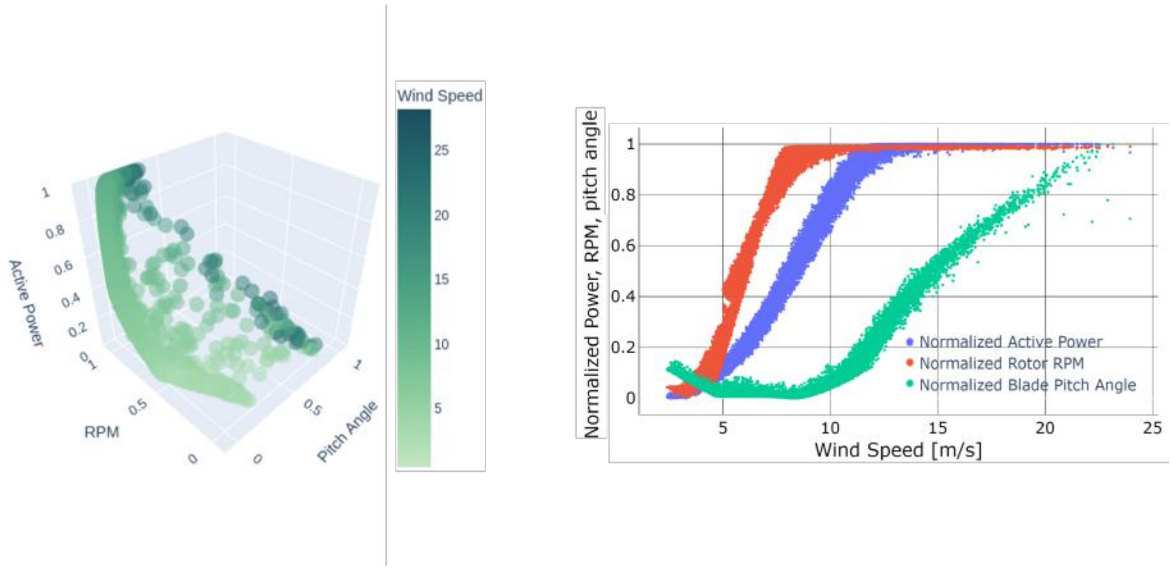


Figure 2 Hyper-surface of target variables for a turbine in farm F0 using raw SCADA dataset. 2D representation of the target variables for the same turbine using the filtered dataset.

where the exponent n has a value ranging from 1 to 3. The term E_{ij} represents the difference in the environmental CPPS between the two turbines, while the terms A_{ij} and C_{ij} represent the differences in scores and the number of predictors obtained by the regression using the the wind turbine variables from raw and filtered datasets. E_{ij} is defined as:

$$E_{ij} = 1 - |CPPS_{env_i} - CPPS_{env_j}| \tag{4}$$

A_{ij} is defined as the difference between two turbines in terms of the weight that data far from characteristic operating conditions has in the calculation of the CPPS:

$$A_{ij} = 1 - |A_i - A_j| \tag{5}$$

where:

$$A_i = 1 - \left| CPPS_{WT_i}{}^{raw} - CPPS_{WT_i}{}^{filt} \right| \tag{6}$$

Likewise, C_{ij} represents the impact that the removal of undesirable operating conditions has on the number of predictors:

$$c_{ij} = \frac{N_i + N_j}{D_i + D_j} \tag{7}$$

In which N_i represents the cardinality of the set obtained as the intersection of the sets of predictors used for the calculation of $CPPS_{WT_i}{}^{raw}$ and $CPPS_{WT_i}{}^{filt}$ and D_i the sum of the cardinality of these two sets of predictors.

Community detection

Network analysis featuring graphs, have been widely used to represent and analyze a large variety of systems, from social networks Scott (1988); Freeman (2004) to time series Bonacina et al. (2020, 2019), in view to unveil communities (of nodes) playing similar roles in the network connected by high concentrations of edges (relations) Rice (1927); Weiss and Jacobson (1955); Homans (2017). Diverse community detection methods have been proposed to date. Traditional methods based on graph partitioning Kernighan and Lin (1970); Barnes (1981), which selects groups of predefined size by minimizing the number of inter-group edges; distance-based methods Lloyd (1982), where a distance function is minimized starting from local network measures. In addition, the spectral algorithms Donath and Hoffman (1973), which detect communities by using the eigenvectors of matrices such as the Laplacian matrix of the graph, and the statistical inference algorithms Hastings (2006); Newman and Leicht (2007).

For the present work, the Louvain's community detection method Blondel et al. (2008) has been used to explore the turbine fleet similarity matrix and to extract communities from large networks.

Specifically, Louvain's method aims at maximizing the modularity of a network Q (representing turbine fleet behaviour stability) defined as:

$$Q = \frac{1}{2m} \sum \left[MPPS_{i,j} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \ , \tag{8}$$

Table 1 Parameters monitored by the wind turbine SCADA system

Variable ID	Description
Active Power	Total active power
I2	Current
Wind Speed	Wind speed
Ambient Temp	Ambient temperature
Spinner Temp	Spinner temperature
Controller Hub Temp	Controller hub temperature
Bearing Temp	HSS gearbox bearing temperature
Gearbox Temp	Gearbox oil temperature
BusBar Temp	Busbar temperature
Generator Phasel, 2, 3 Temp	Generator temperature in stator windings phase 1, 2 and 3
Generator Temp	NDE generator bearing temperature
Generator RPM	Generator RPM
Rotor RPM	Rotor RPM
HVTrafo Phase1, 2, 3 Temp	Temperature in HV transformer phase 1, 2 and 3
Inverter Phase1 Temp	Inverter phase 1 temperature
Rotor Inv Phase1 2, 3 Temp	Rotor inverter phase 1, 2 and 3 temperatures
Blade Pitch Angle	Blade pitch angle
Controller Top Temp	Temperature in the top nacelle controller
Generator SlipRing Temp	Temperature in the split ring chamber
Grid Choke Temp	Grid choke temperature
Hydraulic Pres	Hydraulic pressure
Hydraulic Temp	Hydraulic temperature
Nacelle Temp	Temperature in nacelle
V2	Voltage
Nacelle Dir	Nacelle direction
Wind Dir	Wind direction

where $MPPS_{i,j}$ is the turbine fleet behaviour representation matrix which represents the weight of the edge between node i and node j, k_i and k_j are the sum of the weights of the edges attached to nodes i and j, m is the sum of all the weights in the network, c_i and c_j are the communities of nodes i and j respectively, and δ is the Kronecker delta.

The modularity maximization is obtained with an iterative process: first each node in the network is assigned to its own community and then, for each node i change in modularity ΔQ is calculated when removing i from its own community and moving it into each neighbor j communities.

Among the force-directed methods, the Fruchterman–Reingold algorithm Fruchterman et al. (1991) is adopted, which considers nodes as particles with inertia and edges as springs connecting the particles, and it minimizes the energy content of the dynamic system equivalent to the network.

CASE STUDY

The proposed method is applied to a fleet of 67 multi-megawatt horizontal-axis wind turbines, all of the same model, distributed across five wind farms classified as class 2 under the IEC 61400 standard and situated within a 100 square kilometer area in the Central-Southern Apennines.

All wind turbines are equipped with a SCADA system for monitoring multiple parameters collected from the main components along with ambient measurements. The data, collected every 10 minutes, refers to one year of operation and is composed of 36 monitored variables, including gearbox, generator, and hydraulic system temperatures, as well as active power and wind conditions, as shown in Table 1. Due to its proprietary nature, data cannot be publicly disclosed.

RESULTS AND DISUCUSSION

Figure 3 shows the matrices previously defined, obtained by applying the proposed method to the SCADA data described earlier. Rows and columns of the matrices index a single turbine in the fleet, from 1 to 67, following the indexing method used in Barnabei et al. (2023). Each element of the matrices represents the scores previously defined. The first matrix from the left represents the differences in scores comparing the raw and the filtered dataset. This matrix reveals that the effect of data preprocessing on the predictive capability of the regressor is relatively similar across the entire turbine fleet. Indeed, a minimum correlation value of 0.84 is achieved.

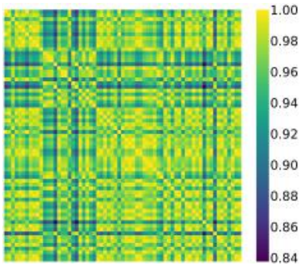
On the other hand, the removal of samples which do not belong to the characteristic curve causes quite different effects

Table 2 Summary of turbines communities characteristics

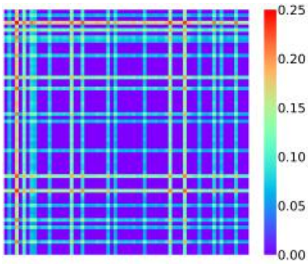
	Turbines number					Median WS		Median P_{norm}		Median RPM		Median h_{active}		Median CF	
	F_1	$F2$	$F3$	$F4$	$F5$	Raw	Filt	Raw	Filt	Raw	Filt	Raw	Filt	Raw	Filt
C 0	5	7	3	1	0	4.37	5.61	0.06	0.15	8.98	10.86	6340.5	3721.5	0.17	0.24
C 1	5	2	4	0	14	6.18	5.19	0.08	0.2	9.27	11.8	7193	4608	0.22	0.33
C 2	0	3	2	20	1	5.18	7.11	0.1	0.3	11.12	13.35	7391.5	4496	0.27	0.42

on the nature and number of predictors: matrix C_{ij}, in fact, shows significantly lower values, with a maximum of 0.25. Matrix E_{ij}, calculated based on the score differences of the environmental CPPS, does not show correlation values greater than 0.6. The similarity matrix, obtained as a composition of the previous matrices and displayed in the figure at the far left, leads, by the method explained in the section dedicated to Community Detection, to the subdivision of turbines into communities.

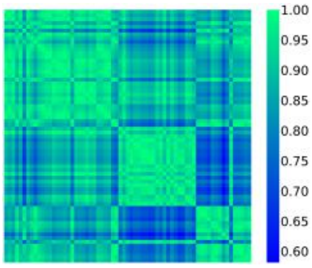
From Figure 3, it is possible to see how the biggest contribution to the similarity matrix is given by matrix E_{ij}. In fact, the large number of zeroes in matrix C_{ij} penalizes the scores coming from matrix A_{ij}, thus only a small amount of entries are increased by the latter.



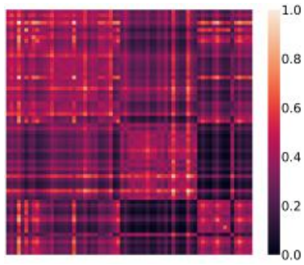
(a) A_{ij}



(b) C_{ij}



(c) E_{ij}



(d) MPPS

Figure 3 Matrices representing the components of the score MPPS (a-c), and the resulting similarity matrix obtained with the composition of the matrices A, C and E (d).

Figure 4 provides a qualitative representation of the obtained communities. The proposed method identifies different behavioral groups entirely derived by the feature selection algorithm applied to community detection. The intersection and union of the sets of most important features for the regression of the target hypersurface, combined with their scores, defined a new set of relations among turbines.

Three distinct groups of turbines are identified: community 0 (C_0) represented in blue, community 1 (C_1) in red, and community 2 (C_2) in green. The communities include wind turbines placed in different wind farms, highlighting the ability of MPPS to establish a relationship across different locations, being invariant to spatial location changes. The main characteristics of each group are reported in Table 2 in terms of median values per community.

Specifically, the median values for wind speed (WS, expressed in m/s), normalized power (P_{norm}, normalized with respect to the nominal power), the number of hours (h_{active}) in which the turbines produced power ($P > 0$), and the capacity factor (CF) obtained from both the raw dataset and the operational characteristics dataset are shown. Additionally, the table presents the number of turbines belonging to each farm per community, referred to as F_i with $i = 1...5$.

Table 3 shows the values of the different average predictive scores obtained for each community. Such values highlight significant variation in $CPPS_{env}$ among the three communities resulting in the high impact that the matrix E_{ij} has on the MPPS. On the other hand, $CPPS_{WT}^{raw}$ and $CPPS_{WT}^{filt}$ assume very similar average values among communites.

Table 3 Summary of CPPS by community

	$CPPS_{env}$	$CPPS_{WT}^{raw}$	$CPPS_{WT}^{filt}$
C 0	0.77	0.92	0.7
C 1	0.65	0.9	0.71
C 2	0.87	0.93	0.71

To provide additional insights into the communities and their emerging different behaviors, we calculate the Probability Density Functions (PDFs) of various variables for each community. For this purpose, we use the non-parametric Kernel Density Estimation (KDE) method with a Gaussian Kernel, as proposed in Scott (2015). The median values of the PDFs per community of different quantities are represented by a continuous line. The bands surrounding the continuous lines are

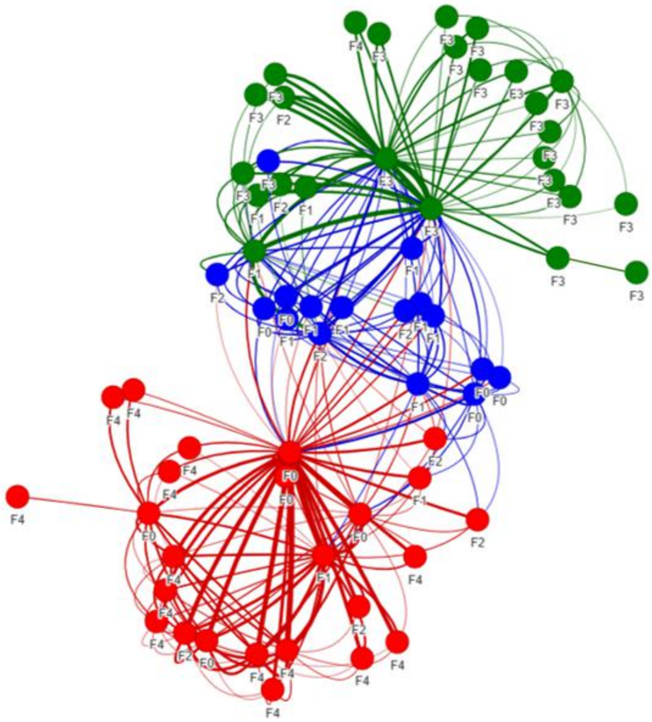


Figure 4 Graph representation of wind turbines fleet communities

evaluated with the standard deviation around the median distribution resulting from the grouping of the PDF related to each turbine in that community.

Figure 5 displays the PDF of the active power (normalized with respect to the nominal power) across the three communities. These communities are clearly differentiated, thereby revealing distinct behaviors among the turbine groups. Specifically, as confirmed by results shown in Table 3, turbines belonging to C_2 are characterized by a higher frequency of operating conditions close to nominal.

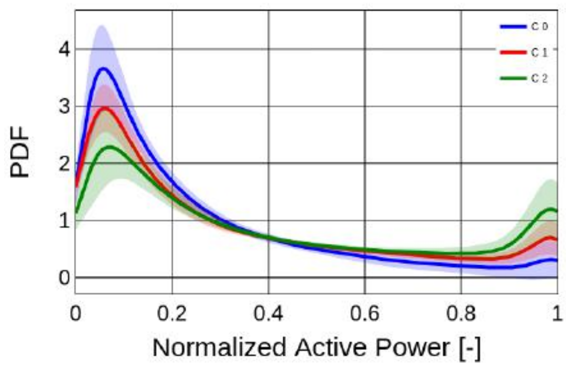


Figure 5 PDFs of normalized active power for each community

Figure 6 displays the PDF of wind speed and turbulence intensity. The frequency at which the turbines in C_0 experience wind speeds at cut-in is higher compared to C_1 and C_2. When observing the nominal wind speed region, C_2 shows a higher occurrence of nominal wind speed, followed by C_1, and finally by C_0.

The turbulence intensity distributions highlight that C_2, characterized by the highest occurrences of optimal conditions in terms of wind speed, also has a higher average TI value. C_0 and C_1 are closer in terms of TI distribution, with the former presenting a low median distribution and being less skewed towards the lowest TI values. Therefore, it is possible to state that C_0 is composed of turbines more often exposed to wind speeds within and below the cut-in region, and less often to wind speeds at or above the nominal wind speed, as confirmed by the active power PDFs in Figure 5. The opposite is true for C_2, which also experiences greater TI at the rotor, with C_1 positioned between C_0 and C_2.

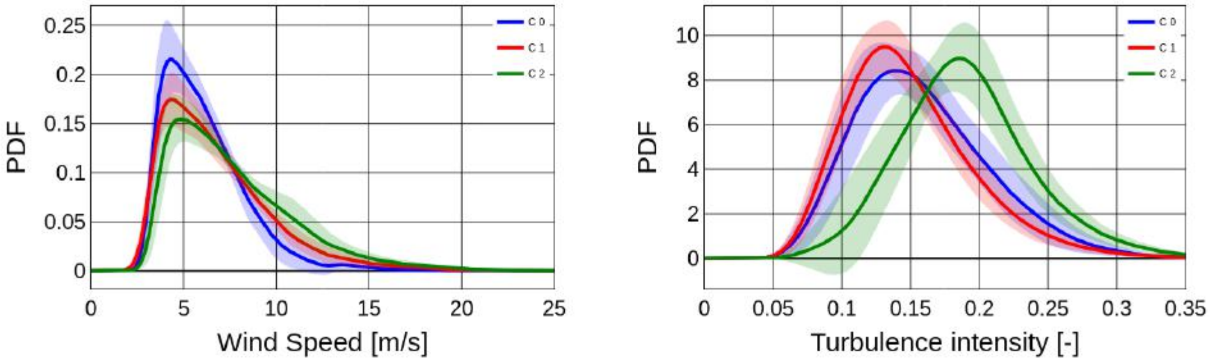


Figure 6 PDFs of wind conditions for each community

Figure 7 shows the PDF of the rotor speed and the tip speed ratio. The latter confirms that C_2 is characterized by a higher occurrence of conditions close to nominal compared to the other two communities. Furthermore, each community has a frequency peak corresponding to the RPM value at the cut-in speed. The position of this peak demonstrates the method's ability to reveal different control strategies: C_2 is characterized by a higher minimum RPM value compared to the other two groups, and this is confirmed by the fact that all the turbines in C_2 (except for one) have a different controller compared to C_0 and C_1.

The PDFs of the tip speed ratio show, for all communities, a peak corresponding to the most frequent wind condition. It can be seen that C_2 has a lower occurrence of this condition at low wind speed and has higher frequencies corresponding to lower TSRs compared to C_0 and C_1.

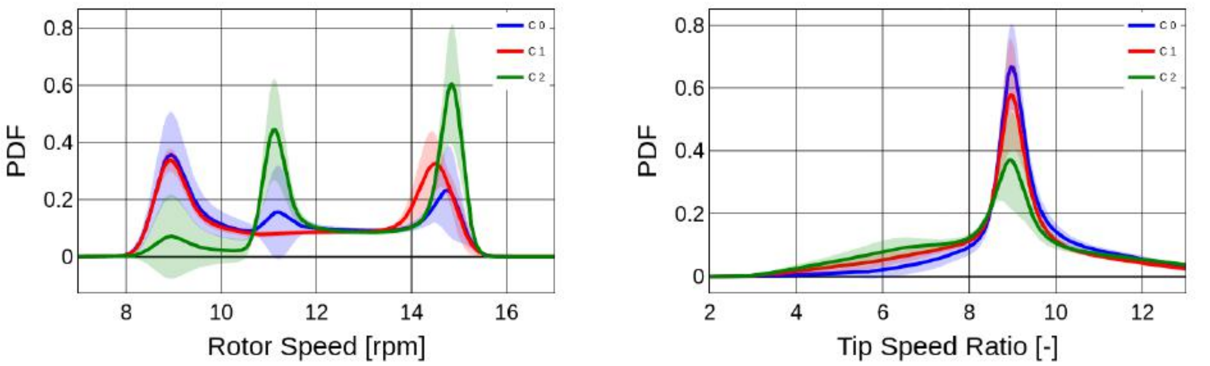


Figure 7 PDFs of RPM and TSR for each community

As an attempt to describe the counterpart of different environmental condition acting on drive-train wind turbine components, Figures 8 and 9 show the PDFs of SCADA signal for some of the most important components of the drive-train, and the pressure and temperature of the hydraulic system. As expected, the different communities emerging from the proposed MPPS score are characterized by different community specific operational state.

Figure 8 displays the PDFs of temperature measurements collected at two different points of the drive train: on the left, the temperatures at the drive end bearings, and on the right, those at the gearbox. C_2, in both cases, exhibit frequency peaks corresponding to higher temperature values compared to those of the other two communities. This is consistent with what observed in the case of the PDFs of active power and wind speeds: C_2 is exposed to higher local wind speed and this results in higher turbine operating regimes and temperatures. C_0 and C_1, on the other hand, present very similar behavior to each other in both cases. Figure 9 shows the PDFs of pressure and temperature of the hydraulic system. On the left, the hydraulic system pressure distribution for the different communities shows that the wind turbines (which are all the same model) present two main operational regimes at two different pressure ranges. C_2, corresponding to the most productive group of turbines, has a less frequent occurrence of low hydraulic pressure and the highest occurrence of high hydraulic pressure among all the communities, with a large spread of the distributions in this region.

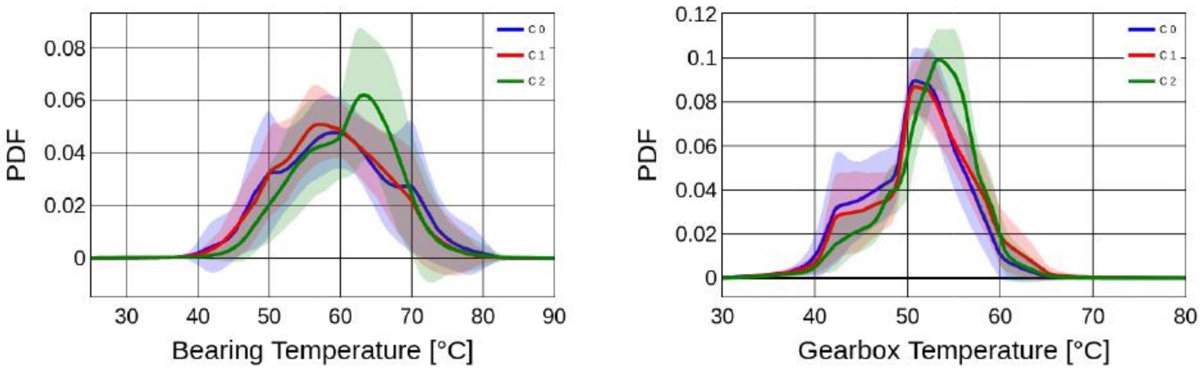


Figure 8 PDFs of temperatures in different locations of the drive-train for each community

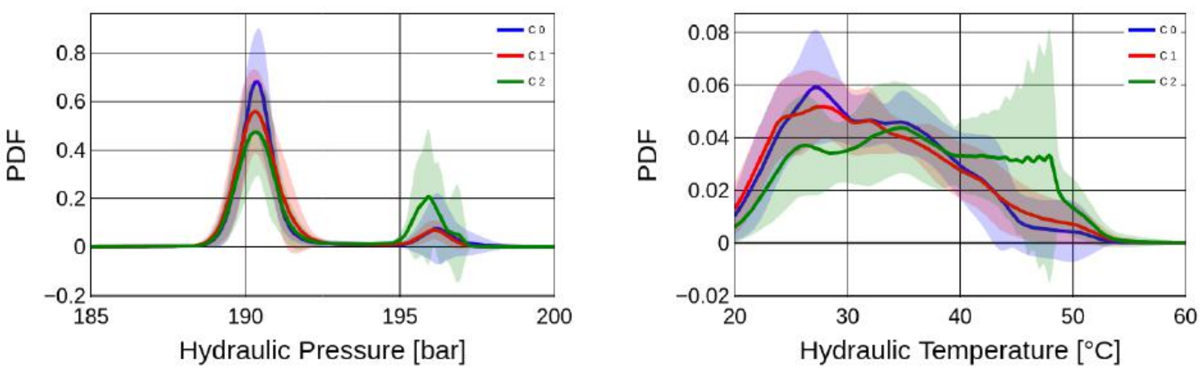


Figure 9 PDFs of pressure and temperature of the hydraulic system for each community

On the right, the distributions of the temperatures of the hydraulic system show that C_0 and C_1 have similar distributions both in terms of median value and spread, with C_0 presenting a slightly higher peak in the frequency for temperatures close to the minimum values. C_2, on the other hand, presents more frequent higher temperatures, with a large spread in the distributions. A possible interpretation for this behavior can be found in the fact that the thermal inertia of the vector fluid in the hydraulic system acts as a sort of memory of the operations that occurred in the recent past, such that a more active wind turbine presents higher temperatures more often than a less productive one.

CONCLUSION

The aim of this work is to extend the use of information provided by the calculation of the CPPS (the score itself as well as the number and nature of the predictors involved) to the HAWT community detection. The method described is applied to a fleet of 67 wind turbines of the same size and model, distributed across 5 different wind farms in the Central-Southern Apennines. The SCADA data from these turbines are pre-processed, yielding two distinct datasets: a raw dataset where no density-based clustering is applied, and a filtered dataset. The CPPS for each turbine is calculated in three cases: considering only environmental variables and considering only wind turbine variables for both the raw dataset and the filtered dataset. From these values, three different comparison matrices for the entire fleet are constructed using an innovative approach: the first based on the calculation of environmental scores; the second and third based, respectively, on the differences between scores and predictors of wind turbines in the filtered and raw cases. From these matrices, the Meta Predictive Power Score (MPPS) is defined, and the fleet similarity matrix is constructed. Results show a significant impact of environmental variables on this matrix.

Using the Louvain community detection method, three different communities (i.e., C_0, C_1, and C_2) of turbines are distinguished. For these communities, the median values of the following variables are evaluated: wind speed, active power, RPM, actual operating hours, and capacity factor. Results show that C_2 consists of turbines subjected to more favorable wind conditions compared to the other two communities. This analysis is confirmed by the evaluation of the probability distributions of wind speed obtained through Kernel Density Estimation. C_2, indeed, is defined by higher occurrences of speeds close to the nominal one compared to C_1 and C_0. The latter, on the other hand, is characterized by a high frequency of occurrences of velocities lower than the cut-in wind speed. Furthermore, the PDF of environmental turbulence is analyzed. C_2 presents a peak frequency at higher turbulence values compared to the other two communities.

Subsequently, focus is placed on the analysis of the PDFs of the temperature values of the main drive-train components (bearings and generator) and the pressure and temperature of the hydraulic system. The results obtained are consistent with those observed for the environmental variables and active power: C_2, composed of the turbines with the highest producibility, also shows higher occurrences of elevated temperatures and pressures.

The communities derived using MPPS are not only defined in terms of operations and producibility (i.e., produced power, wind speed, hours of production, or total number of rotations) but also in terms of asset life consumption, fatigue, and thermal stress that the parts are forced to withstand. Thus, by creating and analyzing the community, it is possible to derive a custom strategy for each community, differentiating any O&M procedure even within the same farm, but sharing the procedure between different communities of the same operator HAWT fleet. The unsupervised nature of this method enables it to automatically identify the unique dynamics that distinguish the operational history and environmental conditions of each turbine from the SCADA data. The proposed method is designed to differentiate between the influences of environmental conditions and turbine-specific operating conditions.

Finally, the methodology is robust against common data management issues in SCADA-based systems, such as inconsistencies in monitored variables across heterogeneous SCADA datasets. This is due to the self-referring and self-contained nature of the derived metrics for each wind turbine when new features are derived in terms of ratios between real or raw operating conditions and ideal or filtered operating conditions.

In spite of the robustness in method design with respect to this issue, further validation on a new heterogeneous dataset is required to assess this feature of the proposed method. Finally, the specific real-time usage of the method is expected to demand case-specific supervision, to better suit the different needs of the O&M operator.

ACKNOWLEDGMENTS

This research is supported by the Ministry of University and Research (MUR) as part of the European Union program NextGenerationEU, PNRR - M4C2 - PE0000021 ”NEST – Network 4 Energy Sustainable Transition” in Spoke 2 ”Energy Harvesting and Off-shore renewables”

REFERENCES

An, E. (2020), ‘Strategy to harness the potential of offshore renewable energy for a climate neutral future’, *European Commission: Brussels, Belgium* .

Barnabei, V., Morvillo, E., Corsini, A. and Bonacina, F. (2023), A systematic study on the use of machine learnt feature derivation in horizontal axis wind turbine fleets, *in* ‘Offshore Mediterranean Conference and Exhibition’, OMC, pp. OMC–2023.

Barnes, E. R. (1981), An algorithm for partitioning the nodes of a graph, *in* ‘1981 20th IEEE Conference on Decision and Control including the Symposium on Adaptive Processes’, IEEE, pp. 303–304.

Blondel, V. D. et al. (2008), ‘Fast unfolding of communities in large networks’, *Journal of statistical mechanics: theory and experiment* **2008**(10), P10008.

Bonacina, F., Corsini, A., Cardillo, L. and Lucchetta, F. (2019), ‘Complex network analysis of photovoltaic plant operations and failure modes’, *Energies* **12**(10), 1995.

Bonacina, F., Miele, E. S. and Corsini, A. (2020), ‘Time series clustering: a complex network-based approach for feature selection in multi-sensor data’, *Modelling* **1**(1), 1–21.

Bonacina, F. et al. (2022), On the use of artificial intelligence for condition monitoring in horizontal-axis wind turbines, *in* ‘IOP Conference Series: Earth and Environmental Science’, Vol. 1073, IOP Publishing, p. 012005.

Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z. and Wagner, D. (2007), ‘On modularity clustering’, *IEEE transactions on knowledge and data engineering* **20**(2), 172–188.

Chen, H., Xie, C., Dai, J., Cen, E. and Li, J. (2021), ‘Scada data-based working condition classification for condition assessment of wind turbine main transmission system’, *Energies* **14**(21), 7043.

Donath, W. E. and Hoffman, A. J. (1973), ‘Lower bounds for the partitioning of graphs’, *IBM Journal of Research and Development* **17**(5), 420–425.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al. (1996), A density-based algorithm for discovering clusters in large spatial databases with noise, *in* ‘kdd’, Vol. 96, pp. 226–231.

Freeman, L. (2004), ‘The development of social network analysis’, *A Study in the Sociology of Science* **1**(687), 159–167.

Fruchterman, T. M. et al. (1991), ‘Graph drawing by force-directed placement’, *Software: Practice and experience* **21**(11), 1129–1164.

Hastings, M. B. (2006), ‘Community detection as an inference problem’, *Physical Review E* **74**(3), 035102.

Homans, G. C. (2017), *The human group*, Routledge.

Ibrahim, H., Ghandour, M., Dimitrova, M., Ilinca, A. and Perron, J. (2011), ‘Integration of wind energy into electricity systems: Technical challenges and actual solutions’, *Energy Procedia* **6**, 815–824.

Izquierdo, J., Crespo Márquez, A., Uribetxebarria, J. and Erguido, A. (2019), ‘Framework for managing maintenance of wind farms based on a clustering approach and dynamic opportunistic maintenance’, *Energies* **12**(11), 2036.

Jin, X., Pan, H., Ying, C., Kong, Z., Xu, Z. and Zhang, B. (2022), ‘Condition monitoring of wind turbine generator based on transfer learning and one-class classifier’, *IEEE Sensors Journal* **22**(24), 24130–24139.

Jin, X., Xu, Z. and Qiao, W. (2020), ‘Condition monitoring of wind turbine generators using scada data analysis’, *IEEE Transactions on Sustainable Energy* **12**(1), 202–210.

Kernighan, B. W. and Lin, S. (1970), ‘An efficient heuristic procedure for partitioning graphs’, *The Bell system technical journal* **49**(2), 291–307.

Kong, X. et al. (2021), ‘Wind turbine bearing incipient fault diagnosis based on adaptive exponential wavelet threshold function with improved cpso’, *Ieee Access* **9**, 122457–122473.

Lloyd, S. (1982), ‘Least squares quantization in pcm’, *IEEE transactions on information theory* **28**(2), 129–137.

Long, H., Sang, L., Wu, Z. and Gu, W. (2019), ‘Image-based abnormal data detection and cleaning algorithm via wind power curve’, *IEEE Transactions on Sustainable Energy* **11**(2), 938–946.

Meyer, A. (2021), ‘Multi-target normal behaviour models for wind farm condition monitoring’, *Applied Energy* **300**, 117342.

Miele, E. S. et al. (2022), Unsupervised feature selection of multi-sensor scada data in horizontal axis wind turbine condition monitoring, *in* ‘Turbo Expo: Power for Land, Sea, and Air’, Vol. 86137, American Society of Mechanical Engineers, p. V011T38A015.

Morrison, R., Liu, X. and Lin, Z. (2022), ‘Anomaly detection in wind turbine scada data for power curve cleaning’, *Renewable Energy* **184**, 473–486.

Newman, M. E. and Leicht, E. A. (2007), ‘Mixture models and exploratory analysis in networks’, *Proceedings of the National Academy of Sciences* **104**(23), 9564–9569.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011), ‘Scikit-learn: Machine learning in python’, *the Journal of machine Learning research* **12**, 2825–2830.

Qiao, F., Ma, Y., Ma, L., Chen, S., Yang, H. and Ma, P. (2021), Research on scada data preprocessing method of wind turbine, *in* ‘2021 6th Asia Conference on Power and Electrical Engineering (ACPEE)’, IEEE, pp. 1606–1610.

Rice, S. A. (1927), ‘The identification of blocs in small political bodies’, *American Political Science Review* **21**(3), 619–627.

Scott, D. W. (2015), *Multivariate density estimation: theory, practice, and visualization*, John Wiley & Sons.

Scott, J. (1988), ‘Trend report social network analysis’, *Sociology* pp. 109–127.

Wang, Y., Infield, D. G., Stephen, B. and Galloway, S. J. (2014), ‘Copula-based model for wind turbine power curve outlier rejection’, *Wind Energy* **17**(11), 1677–1688.

Weiss, R. S. and Jacobson, E. (1955), ‘A method for the analysis of the structure of complex organizations’, *American Sociological Review* **20**(6), 661–668.

Wen, X. and Xie, M. (2021), ‘Performance evaluation of wind turbines based on scada data’, *Wind Engineering* **45**(5), 1243–1255.

Wetschoreck, F., Krabel, T. and Krishnamurthy, S. (2020), ‘8080labs/ppscore: zenodo release’, *Zenodo: London, UK* .

Zhao, Y., Ye, L., Wang, W., Sun, H., Ju, Y. and Tang, Y. (2017), ‘Data-driven correction approach to refine power curve of wind farm under wind curtailment’, *IEEE Transactions on Sustainable Energy* **9**(1), 95–105.