Made available by Hasselt University Library in https://documentserver.uhasselt.be

Measuring wind turbine health using fuzzy-concept-based drifting models

Peer-reviewed author version

Jastrzebska, Agnieszka; MORALES HERNANDEZ, Alejandro; NAPOLES RUIZ, Gonzalo; Salgueiro, Yamisleydi & VANHOOF, Koen (2022) Measuring wind turbine health using fuzzy-concept-based drifting models. In: RENEWABLE ENERGY, 190 , p. 730 -740.

DOI: 10.1016/j.renene.2022.03.116
Handle: http://hdl.handle.net/1942/37485

Measuring Wind Turbine Health Using Fuzzy-Concept-Based Drifting Models

Agnieszka Jastrzebska[a,*], Alejandro Morales Hernández[b], Gonzalo Nápoles[c], Yamisleydi Salgueiro[d], Koen Vanhoof[b]

[a]*Faculty of Mathematics and Information Science, Warsaw University of Technology, Poland.*
[b]*Faculty of Business Economics, Hasselt University, Belgium.*
[c]*Department of Cognitive Science & Artificial Intelligence, Tilburg University, The Netherlands.*
[d]*Department of Computer Sciences, Faculty of Engineering, Universidad de Talca, Campus Curicó, Chile.*

Abstract

Time series processing is an essential aspect of wind turbine health monitoring. In this paper, we propose two new approaches for analyzing wind turbine health. Both methods are based on abstract concepts, implemented using fuzzy sets, which allow aggregating and summarizing the underlying raw data in terms of relative low, moderate, and high power production. By observing a change in concepts, we infer the difference in a turbine's health. The first method evaluates the decrease or increase in relatively high and low power production. This task is performed using a regression model. The second method evaluates the overall drift of extracted concepts. A significant drift indicates that the power production process undergoes fluctuations in time. Concepts are labeled using linguistic labels, which makes our model easier to interpret. We applied the proposed approach to publicly available

*Corresponding author
Email address: A.Jastrzebska@mini.pw.edu.pl (Agnieszka Jastrzebska)

data describing four wind turbines, while exploring different external conditions (wind speed and temperature). The simulation results have shown that turbines with IDs T07 and T06 degraded the most. Moreover, the deterioration was clearer when we analyzed data concerning relatively low atmospheric temperature and relatively high wind speed.
Keywords: time series, concept-based model, regression, wind turbine, health index

---

1. Introduction

Operations and maintenance of wind turbines often benefit from machine learning models devoted to helping domain experts make data-driven decisions. Assessing machine performance degradation and predicting component failure allow planning operations and maintenance costs realistically [1]. This aligns with reliable risk management, which is a crucial aspect in the renewable energy domain [2, 3].

Data-driven wind turbine health estimation methods have gained attention in recent years. On the one hand, the availability of sensor data to construct such models continues to increase [4]. On the other hand, developed countries emphasize moving quickly in tackling climate change, which puts pressure on developing efficient yet human-centric approaches to the management and planning of wind farm operations [5].

The prevailing methodology for health index construction is to build a time series forecasting model and observe the discrepancies between observed and predicted values. Those approaches are model-based since they rely on a secondary model to compute the health index (for example, Zhang

et al. [6] and Yang et al. [7] follow this scheme). Their drawback is that the model fitting stage can be unsuccessful, especially when the data is noisy. A review by Forbes and Zampelli [8] concerning forecasting quality has shown that typically "wind energy forecast errors are not ≪white noise≫, but are characterized by systematic time-series components." This indicates that there is still room for improvement in this domain.

Another group of state-of-the-art approaches requires records concerning several wind turbines and estimates the health of the turbines in a data set, like the method of Liu et al. [9]. This approach works well if we assume some similarity between the historical data and the new (unseen) data. However, the estimator may not generalize well to unseen data if we do not have enough faulty observations to train it, as it was stated by Kusiak and Li [10]. What is more, some of the current methods require vibration data [11] or pictures of blades [12] alongside more commonly used data such as signals from the met mast and produced power.

To overcome the lack of universal methods for measuring wind turbine efficiency deterioration, this paper presents a conceptually distinct approach based on granular computing. More explicitly, we propose two new methods based on fuzzy concepts to explore aging in wind turbines. Concepts can be understood as abstract information granules that summarize underlying raw data. The first method produces a concept-based description of low and high power production. This method extracts a collection of concepts in several windows to capture the time flow and builds concept membership regression models to characterize aging. The second method measures the drift in all concepts observed in time such that a significant drift indicates

(an unwanted) shift in the power production process.

The proposed methods differ significantly from existing approaches since they do not rely on a prediction model but information granules. In that way, the degradation of a wind turbine can be measured from the relative shifts in the data. We recommend executing performance degradation evaluation separately for several different environmental conditions for our methods to be most informative. In this study, we compute scores separately for different wind and temperature conditions for a given wind turbine to study the relation between the environment and the turbine's performance. Our methods are quite universal since they are suitable to evaluate one or more wind turbines while using basic input data found in most physical systems. Moreover, they allow comparing the results for different turbines using both relative and absolute units.

Overall, the rationale of our contribution is to rely on fuzzy concepts to illustrate and estimate the degradation of a wind turbine based on the values of power produced at different operating conditions and granularity levels. Moreover, the fuzzy granulation approach is a suitable formalism to deal with uncertainty in the data. In other words, concepts summarize the information available in raw data and provide a representation robust to noise. As far as we know, the use of fuzzy concepts to measure their drift in time has not been reported in the literature concerning wind turbines.

When it comes to interpretability, the outputs produced by our models are easily comprehended by a human being not acquainted well with the applied machine learning methods. Furthermore, these information granules are divided into groups, and each group has its linguistic label (such as "low

power production" or "high power production"). Such linguistic labels allow deriving comprehensible explanations concerning the operation modes.

Aiming to illustrate our proposal's reliability, we adopt a case study involving four wind turbines. The data is shared publicly by the EDP (*Energias de Portugal*) group under `https://opendata.edp.com/`. The results revealed that analyzing wind tower operations performance for various wind speeds and air temperature conditions is very useful. We observe patterns in the scale of deterioration depending on the environmental conditions. Namely, deterioration in power production was most visible for higher wind speeds.In the analyzed data set, turbines with IDs T07 and T06 deteriorated to a greater extent than turbines T01 and T11.

The remainder of the paper is structured as follows. Section 2 discusses relevant health index measures. Section 3 presents the inner workings of the new method based on fuzzy information granules. Section 4 addresses a case study focused on the analysis of wind turbine data. This is done through a case study involving four wind turbines. Section 5 concludes the paper and provides future research directions.

2. Literature review

Achieving a sufficiently high profit level from generating electric power using wind requires continuous monitoring and systematic maintenance of wind turbines [13]. The former task, wind turbine health monitoring, is often performed with data coming from a SCADA (Supervisory Control And Data Acquisition) system [14, 15]. Such data includes signals from sensors measuring the state of a wind turbine's components and data concerning the

environment such as wind speed, temperature and air pressure.

The literature discusses various approaches to measure wind turbine health, each of them offering distinct predictive capabilities. Zhang et al. [6] presented a method based on a sliding window approach where a third-order polynomial function is fitted to the clean data in each window. Subsequently, the Euclidean distance is used to produce a numerical estimation of turbine health. In contrast, Yang et al. [7] proposed a method for component condition monitoring by considering model prediction residuals. Three statistical indexes (Deviation Index, Volatility Index, and Significance Index) were involved in the computations of the health index. Another conceptually similar approach was presented by Zhan et al. [16]. The authors used Mahalanobis distance between the predicted and actual values to obtain the health index. The same kind of index, comparing predicted and expected values, was discussed in Ren et al. [17]. This health index is based on features obtained with a transfer learning approach.

Liu et al. [9] stressed the necessity of taking into account wind turbine operating conditions. In the cited work, the authors used kernel density estimation to perform a preliminary data split. The health index is expressed in a relatively simple manner. They plot all the data of wind turbines in the group (their method does not work for a single wind turbine) in a single wind speed versus power chart. If for a given wind turbine, some data points deviate from the majority of data from all turbines, they assume it is at a high risk of failure. This method is devoted to wind farms with machines of the same properties nearby. The technique becomes ineffective when this assumption is not fulfilled. Zhang et al. [18] suggest distinguishing four op-

erational condition parameters, which are determined in an unsupervised manner. A health benchmark model was constructed by using the historical data concerning expected turbine operation. The final health index was obtained using the Mahalanobis distance. Other studies also focus on analyzing the operating conditions. For example, Tewolde et al. [19] discussed a method dedicated to offshore wind turbines.

A lively discussion concerns the variables to be used to monitor the turbine. For example, Koukoura [20] mainly used components' temperature to analyze turbine health, while Carroll et al. [11] proposed to include vibration data (not that popular elsewhere) among data from other sensors. Ren et al. [21] use vibration data exclusively. Li et al. [22] advocated for using various SCADA parameters but with an adaptive weight fitted using the analytic hierarchy process. Moreover, there exist approaches that use imaging to analyze wind turbine operations, as presented by Wu et al. [12]. Admittedly, many approaches (e.g., Song et al. [23]) employ a few variables, for instance, wind speed, generated power, and generator speed.

An interesting analysis of logs from three wind turbines with gearboxes in different damage stages was delivered by Lopez et al. [24]. The authors proposed a correlation-based measure to evaluate turbine health. In contrast, Tcherniak [25] focused on rotor health and proposed a method for analyzing structural damages of rotor blades.

Another open problem in the wind turbine health analysis is the lack of methods that focus on user experience. The existing approaches concentrate on producing a numerical estimation of wind turbine health, neglecting the interpretation of the produced score [26]. In this paper, we refocus the ap-

proach to health index construction using fuzzy information granules. The proposed methods use linguistic labels and intuitive visualizations that complement a numerical health index. As a result, the outputs produced by our methods are straightforward to interpret by domain experts who are not experts in machine learning.

3. Measuring the health of wind turbines with fuzzy drifting concepts: two methods

This section outlines the details of the new methods for wind turbine health evaluation. More explicitly, these methods measure power production deterioration through changes observed in concepts. Both methods consist of the same set of preliminary steps: (i) Data preprocessing, (ii) Operating conditions binning, and (iii) Concepts extraction. The first method extracts information about relatively low and relatively high power production and then evaluates whether there was an increase in the former and a decrease in the latter that would suggest aging. The second approach measures the discrepancy in power production in time. Figure 1 outlines the procedure to compute the proposed health indexes.

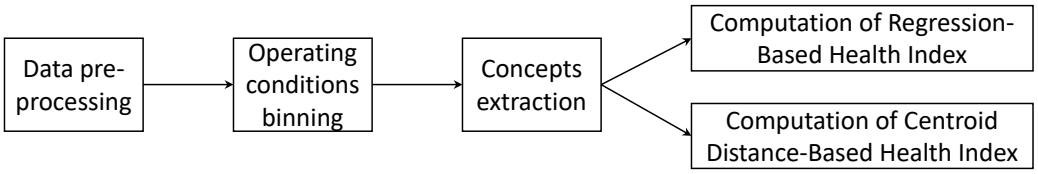


Figure 1: Key steps to compute the proposed health indexes.

The proposed solution serves as a decision-making support tool. We envision that an analyst will run the program we prepared and based on

the output recommend some decisions concerning the physical system. We implemented a script that an analyst can run on a standard PC. The script reads in data stored locally with SCADA signals in a CSV file and outputs tables and plots visualizing subsequent steps of the procedure. The program was written in R with the use of e1071[1] package.

3.1. Data preprocessing

The first step relies on elementary data cleaning to produce active power time series. More explicitly, we remove observations corresponding to very low and very high wind values. To determine what should be the cutting thresholds for low and high wind values, we can use, for example, the turbine's theoretical power curve if available. We can inspect the wind-power plot if we do not wish to use a theoretical power curve. In this plot, we must determine the wind values between which the diagonal part of the data scatter is observed. Narrowing down the data in this way lets us avoid taking into account instances describing the take-off and the saturation of the wind turbine. Equation (1) shows how to compute the ratio between power value and wind value for each moment in time,

$$pw_i = \frac{p_i}{w_i}, i = 1, 2, \ldots, L \tag{1}$$

where L is the length of the time series obtained from the sensors, p_i and w_i denote the power and wind values, respectively, in the i-th time step. Equation (1) can be seen as an analogy to the Tip Speed Ratio which is a ratio of the tip speed of blade to the wind speed, used for instance in [27].

[1] https://cran.r-project.org/web/packages/e1071/index.html

In our approach, we use the ratio from Equation (1) to determine which observations fall between the first and the third quantile of data distribution. Data points that fall into this range are kept, while remaining data points are assumed to be outliers and consequently removed. In our case, we analyzed power generated for wind values between 4.5 and 9 m/s.

3.2. Operating conditions binning

The second step of the procedure aims at splitting the cleaned data into segments (bins), each describing different environmental conditions. We recommend using two time series: wind speed and air temperature. If the air temperature measurements are not available, we can use the air pressure variable as an alternative.

We can use two approaches to determine the range of each bin. A naive method is to use a fixed value for each segment. However, this can be counterproductive since it may lead to empty bins or bins with few observations. An alternative is to use a clustering algorithm to select the length of each bin based on the observations clustered around a center. The k-means algorithm [28] is a suitable choice. The value k corresponds to the number of clusters and, in our case, the number of bins we want to extract.

We analyzed wind speed in bins each of length of 0.5 m/s, as it is the value frequently used in the literature [29]. The temperature was split into intervals determined using a clustering procedure (as it is performed in other studies, including [18]). In our research, we considered four clusters for temperature values.

3.3. Concepts extraction

In this step, we measure wind turbine health for each wind and temperature conditions defined on the extracted bins. Next, we sort the generated power values over time and split this sequence into Q non-overlapping windows. Let us denote the resulting length of the window as R. If the data is collected with a 10-minutes frequency, then R=40 or R=60 would be reasonable configurations. Each window w contains P observations arranged in a sequence of active power values as follows:

$$z_1^{(w)}, z_2^{(w)}, z_3^{(w)}, \ldots, z_P^{(w)}. \tag{2}$$

The superscript (denoted with w) was introduced to emphasize that the sequence defined in (2) concerns a window, not the entire time series. Next, we transform that sequence into a two-dimensional space by computing successive increments of time series values. The i-th data point is represented with a pair $(z_i^{(w)}, dz_i^{(w)})$ denoting the i-th amplitude and the i-th change (increment) of amplitude, where $dz_i^{(w)} = z_i^{(w)} - z_{i-1}^{(w)}$. This means that the active power can be represented as follows:

$$(z_2^{(w)}, dz_2^{(w)}), (z_3^{(w)}, dz_3^{(w)}), \ldots, (z_P^{(w)}, dz_P^{(w)}). \tag{3}$$

We extract concepts using fuzzy c-means clustering [30, 31] for each window (now represented in two-dimensional space). This algorithm extracts centroids representing the discovered clusters while being endowed with a membership function that links data points to each centroid. The number of centroids, denoted as C, has to be set a priori. The clusters are fuzzy

since one data point belongs to more than one cluster simultaneously with a membership value in the $[0, 1]$ interval.

We set C=3 so that we can obtain centroids representing *low*, *moderate*, and *high* power values. Wind turbine wear will manifest with: (i) Memberships to centroids representing high values that diminish with time, and (ii) Memberships to centroids representing low values that increase with time.

The fuzzy c-means algorithm creates a matrix $\mathbf{U} = [\mu_{ij}], \mu_{ij} \in [0, 1], i = 1, \ldots, P-1, j = 1, \ldots, C$. Each element μ_{ij} denotes the degree of membership of the i-th data point to the j-th fuzzy cluster. The procedure involves a fuzzification coefficient $m \in \mathbb{R}$, such that $1.0 < m < \infty$. Large m attenuates memberships values so we can achieve less differentiation in concept belongingness, or in other words, clusters are fuzzier. In contrast, m close to one makes the fuzzy c-means act as a crisp clustering algorithm. In [32], the authors recommend setting m to 2.0.

The fuzzy c-means algorithm iteratively minimizes the following objective function:

$$J_m = \sum_{i=1}^{P-1} \sum_{j=1}^{C} \mu_{ij}^m \|\mathbf{z}_i - \mathbf{v}_j\|^2, \tag{4}$$

where $\mathbf{z}_i$ is the i-th data point, $\mathbf{v}_j$ is the j-th cluster, m is the fuzzification coefficient, μ_{ij} is the degree of membership of $\mathbf{z}_i$ to the j-th cluster, and $\|\cdot\|$ is a norm to measure the distance between a data point and a centroid [33].

The iterative procedure adjusts the membership values μ_{ij} from the matrix $\mathbf{U}$ and the cluster center $\mathbf{v}_j$ as follows:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|\mathbf{z}_i - \mathbf{v}_j\|}{\|\mathbf{z}_i - \mathbf{v}_k\|} \right)^{2/(m-1)}} \tag{5}$$

and

$$\mathbf{v}_j = \frac{\sum_{i=1}^{P-1} \mu_{ij}^m \cdot \mathbf{z}_i}{\sum_{i=1}^{P-1} \mu_{ij}^m}. \tag{6}$$

The procedure terminates when the largest change in $\mathbf{U}$ from the k-th to the $(k+1)$-th iteration falls below a given ε threshold or when a predefined number of iterations has been exceeded [34].

As mentioned, concept extraction is performed on each window. Next, we sort the concepts in the decreasing order (high, moderate, low). After completing this step, for each bin we obtain $Q \cdot C$ concepts labeled with the bin number for which they were extracted.

3.4. Concept analysis and health index approaches

We can use extracted concepts to visualize and evaluate the degradation tendency. We propose two methods to quantify the power production deterioration with time: (i) A regression-based method utilizing drifting concept memberships and (ii) A distance index-based method that measures the degree of concept drift.

3.4.1. Method 1: computation of the regression-based health index

The first method uses a regression analysis using the membership values of data points to clusters representing the "high-power" concept. Concepts

reside in a two-dimensional space of defined by the power value and change of value (see Equation (3)) and each concept is represented with a centroid. We sort the centroids according to the power value (the first dimension) in the decreasing order: the first centroid represents data points with high values of generated power while the third centroid represents underlying data points with low values of generated power.

Let us assume that we process the k-th window (out of Q windows in a given bin) and obtain a sequence with the form $\mu_{1,1}^{(k)}, \mu_{2,1}^{(k)}, \ldots, \mu_{P-1,1}^{(k)}$, $\mu_{1,2}^{(k)}, \mu_{2,2}^{(k)}, \ldots, \mu_{P-1,2}^{(k)}$, $\mu_{1,3}^{(k)}, \mu_{2,3}^{(k)}, \ldots, \mu_{P-1,3}^{(k)}$, where $\mu_{i,j}^{(k)}$ denotes the membership of the i-th data point to the j-th centroid, both in the k-th window. $P-1$ is the number of data points in the two-dimensional space of value and change of value. The end index is $P-1$, because we computed the lags, as in Equation (3). We concatenate the series from each window to obtain the following three ordered sequences:

$$\mu_{1,1}^{(1)}, \mu_{2,1}^{(1)}, \ldots, \mu_{P-1,1}^{(1)}, \mu_{1,1}^{(2)}, \mu_{2,1}^{(2)}, \ldots, \mu_{P-1,1}^{(2)}, \ldots, \mu_{1,1}^{(Q)}, \mu_{2,1}^{(Q)}, \ldots, \mu_{P-1,1}^{(Q)} \tag{7}$$

$$\mu_{1,2}^{(1)}, \mu_{2,2}^{(1)}, \ldots, \mu_{P-1,2}^{(1)}, \mu_{1,2}^{(2)}, \mu_{2,2}^{(2)}, \ldots, \mu_{P-1,2}^{(2)}, \ldots, \mu_{1,2}^{(Q)}, \mu_{2,2}^{(Q)}, \ldots, \mu_{P-1,2}^{(Q)} \tag{8}$$

$$\mu_{1,3}^{(1)}, \mu_{2,3}^{(1)}, \ldots, \mu_{P-1,3}^{(1)}, \mu_{1,3}^{(2)}, \mu_{2,3}^{(2)}, \ldots, \mu_{P-1,3}^{(2)}, \ldots, \mu_{1,3}^{(Q)}, \mu_{2,3}^{(Q)}, \ldots, \mu_{P-1,3}^{(Q)} \tag{9}$$

Equation (7) contains membership values to high power production concepts (centroids), Equation (8) contains membership values to moderate power production concepts, and Equation (9) encloses membership values to low production centroids. Next, we compute a regression model with the form:

$$\mu_{i,1} = a_1 x_i + b_1 + \varepsilon_1. \tag{10}$$

In this regression model, which is fitted with the data given in Equation (7), x denotes the observation order in time. A negative value of the slope a_1 indicates aging since it suggests a decrease in the membership to the centroids corresponding to the high power production. Let us recall that centroids were sorted in a decreasing order; they describe (1) high, (2) moderate, and (3) low power production. This is the reason why we use the subscript 1 in the regression coefficient a_1. We assume that a linear model will be suitable yet straightforward to depict the turbine aging.

We can perform an analogous analysis for the concepts representing low time series values. According to our method, aging would be manifested as an increasing membership to these fuzzy concepts. Aiming at evaluating whether these membership values increase or decrease with time, we can compute a regression model with the following form:

$$\mu_{i,3} = a_3 x_i + b_3 + \varepsilon_3, \tag{11}$$

where the subscript 3 in a_3 indicates that we deal with the third (and last) concept concerning low power production. The regression model given by Equation (11) is fitted using the sequence given in Equation (9). Please note that elements gathered in sequence (8) are not considered in health evaluation.

An important feature of this aging index is that coefficients a_1 and a_3 are comparable for any wind turbine in any data set. Membership values are always in the $[0, 1]$ interval, thus the models are created on normalized data. These values describe relative aging, not in natural units (like watt-hours) that depend on mechanical properties. We are not constructing a regression

model for moderate power production concepts, because the interpretation of the resulting coefficient will not be intuitive.

3.4.2. Method 2: computation of the centroid distance-based health index

The idea behind the second health index is to measure the discrepancy between *high* and *low* power production centroids. It is measured separately in each wind-temperature bin. In each bin, we have a sequence of observations that we split into windows. Let us recall that for each window (we have Q windows ordered in time), we created C clusters to represent the values falling into it. Clusters describe *high*, *moderate*, and *low* power production. However, interpreting $Q \cdot C$ centroids visually may be inconvenient for domain experts. Thus, we propose an additional aggregation option to simplify the interpretation.

Let us formalize this discussion by denoting the extracted centroids as presented in Equations (12), (13), and (14),

$$\mathbf{v}_1^{(1)}, \mathbf{v}_1^{(2)}, \ldots, \mathbf{v}_1^{(Q)}, \tag{12}$$

$$\mathbf{v}_2^{(1)}, \mathbf{v}_2^{(2)}, \ldots, \mathbf{v}_2^{(Q)}, \tag{13}$$

$$\mathbf{v}_3^{(1)}, \mathbf{v}_3^{(2)}, \ldots, \mathbf{v}_3^{(Q)}. \tag{14}$$

Equation (12) covers centroids representing high power production, Equation (13) represents moderate power production, and Equation (14) contains centroids concerning low power production. They are sorted according to the window for which they were created.

We can generalize further the information carried by these $Q \cdot C$ centroids by splitting each type of centroids into two clusters, again using a centroid-

based algorithm. We use the fuzzy c-means method again (for consistency). We forward the sequence from Equation (12) to the input of the fuzzy c-means clustering algorithm. The same is done separately for data given in Equation (13) and Equation (14). Each time we set the desired number of clusters to two. The new centroids do not capture time flow, but we can use them to describe the discrepancy in the data. We extract two centroids for each group, because our goal will be to measure the discrepancy, and for this a pair of centroids is enough. We can sort them and interpret one as a centroid representing low values and the other one as a centroid describing high power production values. Obtained pairs of centroids are denoted as $\mathbf{v}_{1L}, \mathbf{v}_{1H}$, $\mathbf{v}_{2L}, \mathbf{v}_{2H}$, and $\mathbf{v}_{3L}, \mathbf{v}_{3H}$. In each pair, the centroid with letter L corresponds to low power production values, and the centroid with letter H corresponds to high power production values. The intuition of this measure is that the more separated the centroids, the more aging in the turbine. This distance will be zero in an ideal world scenario where there is no change in the performance. Equation (15) shows the distance index (DI) formalizing this turbine health index:

$$DI = \sum_{i=1}^{3} d(\mathbf{v}_{iL}, \mathbf{v}_{iH}). \tag{15}$$

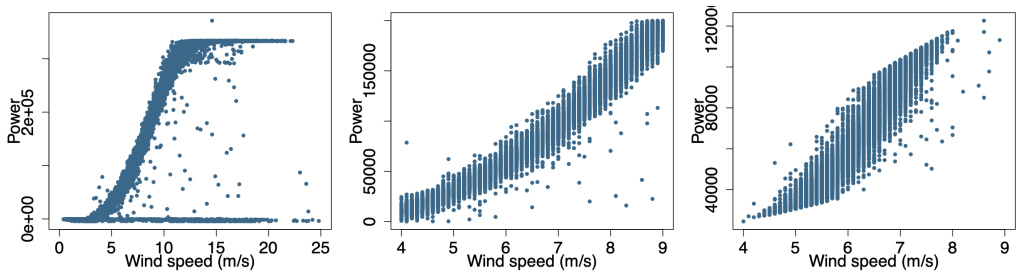
where $d(\cdot, \cdot)$ denotes any distance measure (e.g., the Euclidean distance). We interpret large DI values as a negative phenomenon. Let us recall that the centroids are extracted for fixed wind and air temperature values, so we are justified to demand as small differentiation in machine performance as possible. Notice that the DI health index will provide better insight when computed for more than one wind-temperature bin.

4. Case Study: EDP data set

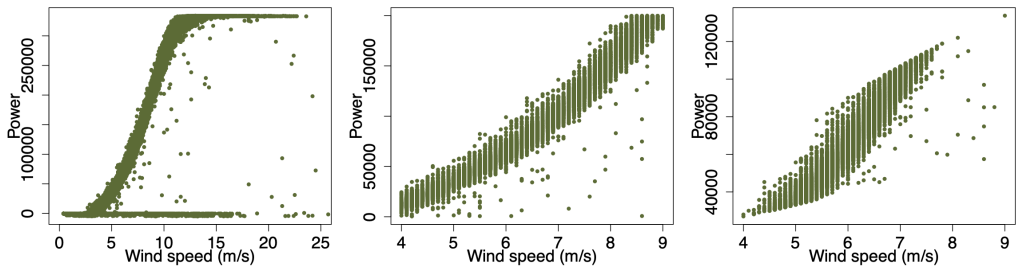
In this section, we present the application of the proposed procedure to the EDP (*Energias de Portugal*) data set available under `https://opendata.edp.com/`. The case study concerns four wind turbines SCADA signals for 2016 and 2017. There were five wind turbines, but we picked randomly four as we believe it is enough to succinctly demonstrate a method for health evaluation. Further parts of this section are split into three subsections. In Subsection 4.1, we present an introductory analysis of the results concerning the turbine with ID T11. We chose to discuss T11 in greater detail because later comparative analysis showed that the deterioration of performance of this turbine was relatively moderate. Subsection 4.2 compares the results for four turbines in the EDP data set and highlights interesting observations. Subsection 4.3 presents the application of the Distance Index for turbine performance evaluation.

Figure 2 shows some statistics concerning all wind turbines' power and wind speed variables. More explicitly, the first column displays plots of the relationship between power and wind speed for each analyzed wind turbine. Data scatter follows the usual sigmoid-like shape. The data is noisy as expected (i.e., we have plenty of outliers and in T01, and there is one anomalous observation). The second column concerns the data after narrowing it down to the wind speed values between 4 and 9 m/s. The third column contains plots after selecting observations between the first and third quantile with the use of the ratio given in Equation (1).

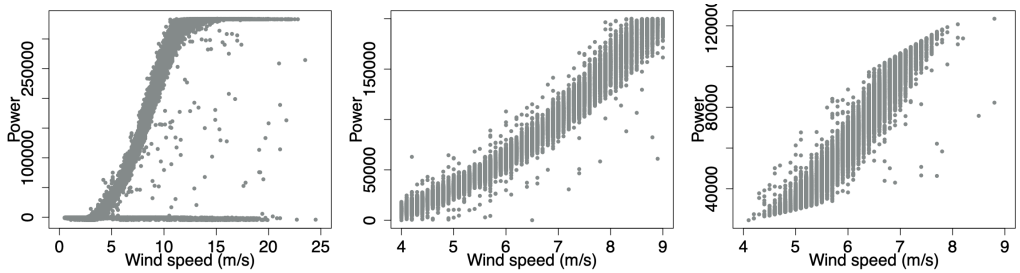
After cleaning, the data is transformed to the two-dimensional space of value and change of value, as shown in Equation (3). The available data



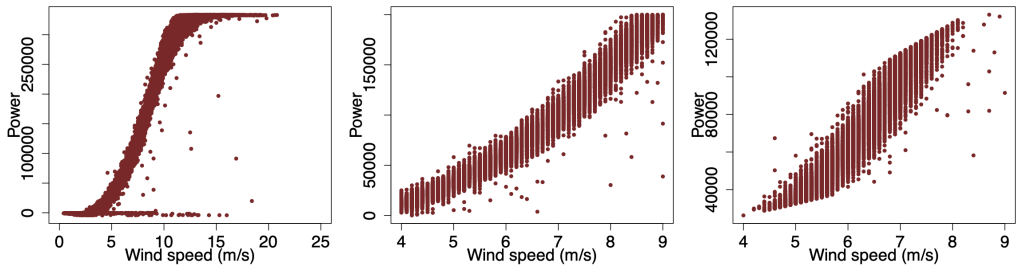
(a) T01



(b) T06



(c) T07



(d) T11

Figure 2: Power versus wind speed for each wind turbine (turbine IDs are given in sub-captions). The first column contains all data points available in the EDP data sets. The second column contains data points after narrowing down the wind range. The third column shows the remaining points after applying the preprocessing step.

points are partitioned into bins, each corresponding to a specific range of wind values. In this case study, bins started at 5 m/s and ended at 8 m/s, which is the range of wind values mainly represented in the data after cleaning, cf. Figure 2, the third column. We consider wind bins of length 0.5. The second environmental factor that we consider is atmospheric temperature. Thus, we split the observations into four smaller bins corresponding to different temperature conditions in each wind bin.

4.1. Study concerning turbine T11

In the case study concerning turbine T11, we use four temperature bins, with centroids at ca. 15, 18, 22, and 27 degrees Celsius. In each bin, we first split the data into $Q = 30$ windows. Then, we execute fuzzy c-means and extract $C = 3$ concepts.

Figure 3 displays concepts (centroids) extracted for different wind ranges for the turbine T11. We have three concepts for each window. Thus, each plot contains $3 \cdot 30$ points. We sorted concepts according to the increasing value of produced power. Points are marked using three symbols: a square (*low* power), a circle (*moderate* power), and a triangle (*high* power). Concepts introduce a natural discretization and summarization of raw data. Concepts in Figure 3 are colored. The yellower the color, the more recent the observation while the navy color represents the oldest observations. In the scatter plots, we observe a systematic shift of colors from the right (high power) to the left (low power) for each wind turbine. Yellowish colors are on the left, and blueish colors tend to be positioned on the right. This behavior is a clear indication of performance deterioration.

Gradient coloring representing the time flow enables visual comparison

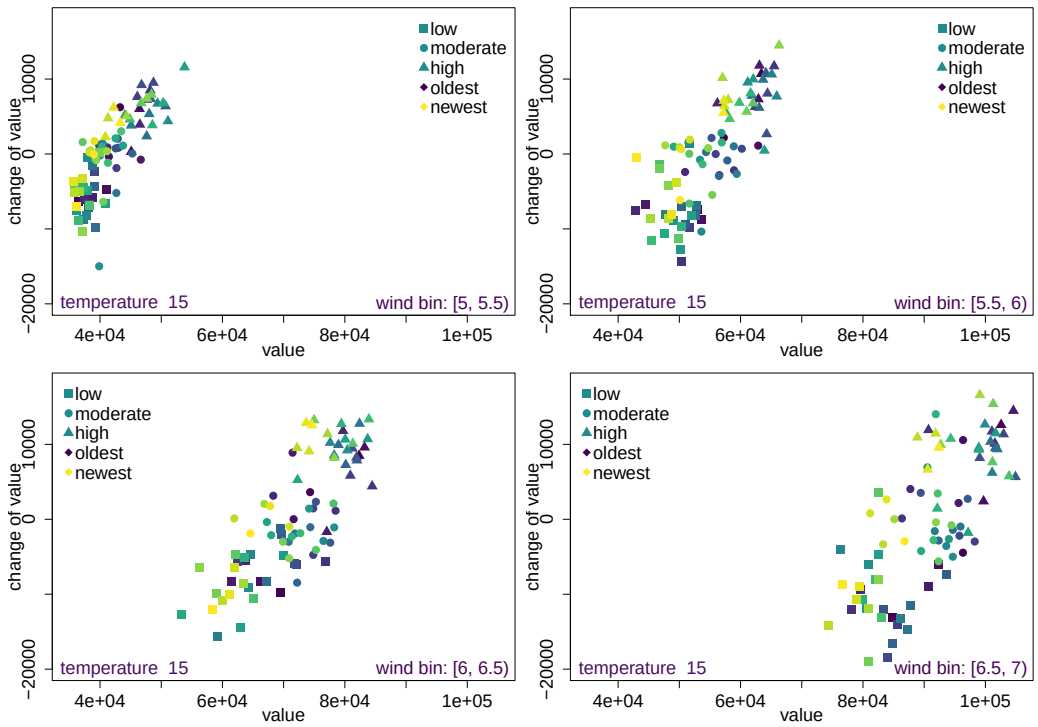


Figure 3: Concepts in the two-dimensional space of power value and change of power value in turbine T11. In each window, three concepts corresponding to *low* (squares), *moderate* (circles), and *high* (triangles) values were extracted. The most recent concepts are colored yellow. It is worth mentioning that the temperature interval is the same in each plot while the wind interval changes.

of the degree of deterioration of power production. Such illustration is easy to understand by domain experts. Splitting the data into bins helps analyze the differences between the machine's operating characteristics under different weather conditions. Concepts and membership to concepts can both be interpreted straightforwardly. Concepts provide a data-driven discretization of the values present in the data. Furthermore, we attach linguistic labels to describe the values they represent. The use of $C = 3$ concepts entails that we

describe the tendency of the power production to be low, moderate, or high. The use of human-friendly linguistic labels "low", "moderate", and "high" make the analyzes convenient.

Subsequently, we present the health index based on regression models. Below, we present analyses for wind ranges from 5 to 7 m/s and four temperature bins. In each wind and temperature bin, we computed a regression model using Equation (10). Such a regression model concerns membership values to high power production concepts. Table 1 presents the obtained regression slopes. Negative values indicate a decreasing trend that entails turbine performance degradation.

Table 1: Regression-based health index for turbine T11. The slopes correspond to the regression models fitted for memberships to the *high* power production concepts. Negative values indicate aging. Models were fitted separately for different wind and temperature conditions. Slopes were multiplied by 10^5 to enhance clarity.

		Wind bins				
		[5, 5.5)	[5.5, 6)	[6, 6.5)	[6.5, 7)	[7, 7.5)
Temp. (°C)	15	−6.64	3.71	5.29	−3.94	−24.76
	18	−6.78	−6.69	−11.79	−6.99	−35.98
	22	−1.34	−18.67	−2.99	−0.31	5.75
	27	0.61	5.62	−12.03	10.09	−1.13
	sum	**−14.15**	**−16.03**	**−21.52**	**−1.15**	**−56.12**

In Table 1, we observe negative slopes in most cases. It is challenging to interpret the results concerning the highest temperatures, around 27°C. In this case, slopes are often positive. When looking at the temperature values histogram in Figure 4 concerning preprocessed data, we notice that these values are infrequent.

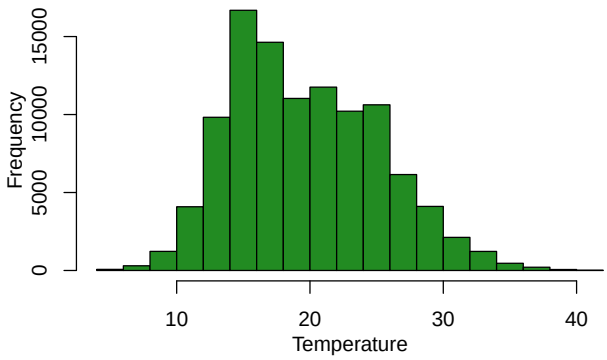


Figure 4: Histogram of temperature values collected at the wind turbine.

Table 2 presents the regression slopes of models computed for membership values to the low power production concept. They were computed according to Equation (11). In this case, positive values indicate that there is an undesired shift, because they mean that the turbine started to output lower power values. In Table 2, in the vast majority of cases (i.e., in the majority of analyzed wind and temperature bins), the result indicates aging because values have a positive sign.

4.2. Comparative analysis

Subsequently, we compare regression model-based health indexes computed for all wind turbines in the EDP data set. This experiment still focuses on four temperature bins around 15, 18, 22, and 27 degrees Celsius. We do not recommend increasing the number of temperature bins since it reduces the number of data points in each bin. In these experiments, we change the parameter Q to 10.

Table 3 presents the slopes of the regression model fitted to membership values concerning concept describing *high* power production for different

Table 2: Regression-based health index for turbine T11. The slopes correspond to the regression models fitted the data under the *low* concept. Positive values indicate aging. Models were fitted separately for different wind and temperature conditions. Slopes were multiplied by 10^5 to enhance clarity

		Wind bins				
		[5, 5.5)	[5.5, 6)	[6, 6.5)	[6.5, 7)	[7, 7.5)
Temp. (°C)	15	−5.29	−1.47	−3.44	4.80	28.54
	18	−4.73	6.96	11.11	−4.26	25.03
	22	0.79	7.94	−0.53	−7.77	−1.34
	27	31.34	2.91	0.47	−7.86	−11.97
	sum	**22.12**	**16.34**	**7.62**	**−15.09**	**40.26**

weather conditions. In many cases, we obtained a negative slope indicating decreasing health of the studied turbines. We may infer that T01 deteriorated the least, while T07 deteriorated the most. On the other hand, looking the slopes obtained for the *low* concept (Table 4), performance deterioration is confirmed with positive sign. For this concept, the largest slopes concern the $[7, 7.5)$ wind range.

In Table 3, we can observe that memberships to the concepts describing high power production are declining with time. This means that high power production decreases with time. In contrast, as illustrated in Table 4, memberships to the concepts describing small power production increase with time (slopes have positive signs). In other words, we observe the degradation of the power production process for fixed environmental conditions. Tables 3 and 4 show that T01 turbine is the least affected by the aging processes, while T06 and T07 are the most affected.

Figure 5 presents plots concerning membership values ordered in time to

Table 3: Regression model slopes (Method 1) computed for membership values to the *high* concept and $Q = 10$. Negative values indicate aging. Models were fitted separately for different wind and temperature conditions. Slopes were multiplied by 10^5 to enhance clarity.

Turbine ID	Temp. (°C)	Wind bins (m/s) [5, 5.5)	[5.5, 6)	[6, 6.5)	[6.5, 7)	[7, 7.5)
T01	15	7.49	3.46	−0.70	1.65	30.14
	18	5.02	−9.39	−23.84	−3.46	−24.70
	22	−11.13	4.86	−11.45	6.60	−23.09
	27	31.82	35.94	22.93	6.16	−69.74
	sum	**33.20**	**34.87**	**−13.06**	**10.95**	**−87.39**
T06	15	−5.16	−1.92	5.33	−2.38	−207.84
	18	−16.24	12.03	9.30	5.55	−93.16
	22	18.62	−15.73	2.52	19.74	−0.97
	27	−18.13	12.08	−3.98	54.72	−138.53
	sum	**-20.91**	**6.46**	**13.17**	**77.63**	**−440.50**
T07	15	−9.75	8.17	−6.15	−7.58	−185.33
	18	−1.62	−12.84	−0.42	2.20	−147.45
	22	−7.46	−4.50	−14.56	−12.69	−114.85
	27	4.33	−26.02	36.28	24.59	−78.23
	sum	**−14.50**	**−35.19**	**15.15**	**6.52**	**−525.86**
T11	15	0.10	9.71	1.07	−5.27	−42.69
	18	−9.17	−15.26	5.81	14.72	−50.96
	22	−4.64	−23.66	−1.52	−16.08	−2.17
	27	10.96	−10.08	−20.96	−83.84	−3.51
	sum	**−2.75**	**−39.29**	**−15.60**	**−90.47**	**−99.33**

visualize the results. The top row concerns memberships to the large power production concepts. The bottom row concerns memberships to the low power production concepts. We have fitted linear regression models in both cases, and the obtained lines are visible in the plots. We adopt the slopes

Table 4: Regression model slopes (Method 1) computed for membership values to the *low* concept and $Q = 10$. Positive values indicate aging. Models were fitted separately for different wind and temperature conditions. Slopes were multiplied by 105to enhance clarity.

Turbine ID	Temp. (°C)	Wind bins (m/s)				
		[5, 5.5)	[5.5, 6)	[6, 6.5)	[6.5, 7)	[7, 7.5)
T01	15	−4.44	9.92	2.28	−5.75	−6.62
	18	−4.15	18.65	15.46	−1.49	0.24
	22	6.58	−9.67	7.56	−6.51	3.42
	27	46.01	2.18	−33.05	20.79	25.07
	sum	**44.00**	**21.08**	**−7.75**	**7.04**	**22.11**
T06	15	−0.64	8.60	−4.45	−10.88	91.83
	18	4.70	−11.88	−0.22	−11.41	65.37
	22	−11.44	26.30	−4.07	−14.05	60.53
	27	0.51	−19.64	−3.60	−40.47	48.70
	sum	**−6.87**	**3.38**	**−12.34**	**−76.81**	**266.43**
T07	15	6.41	−4.31	10.45	−3.39	78.83
	18	0.83	19.26	4.39	16.17	98.68
	22	3.00	1.47	21.98	8.77	61.85
	27	−25.91	23.40	−14.86	−13.95	−176.66
	sum	**−15.67**	**39.82**	**21.96**	**7.60**	**62.70**
T11	15	10.67	−1.83	−8.12	7.52	57.14
	18	−15.90	7.86	3.33	−28.14	56.68
	22	8.10	0.42	−0.02	−4.44	7.21
	27	10.04	−5.51	1.14	53.18	−24.97
	sum	**12.91**	**0.94**	**−3.67**	**28.12**	**96.06**

of the obtained linear models to inspect turbine performance degradation. Negative signs of regression models in the first row indicate aging (they tell that high power production declines with time). In contrast, positive signs in the bottom row indicate aging: low power production increases with time.

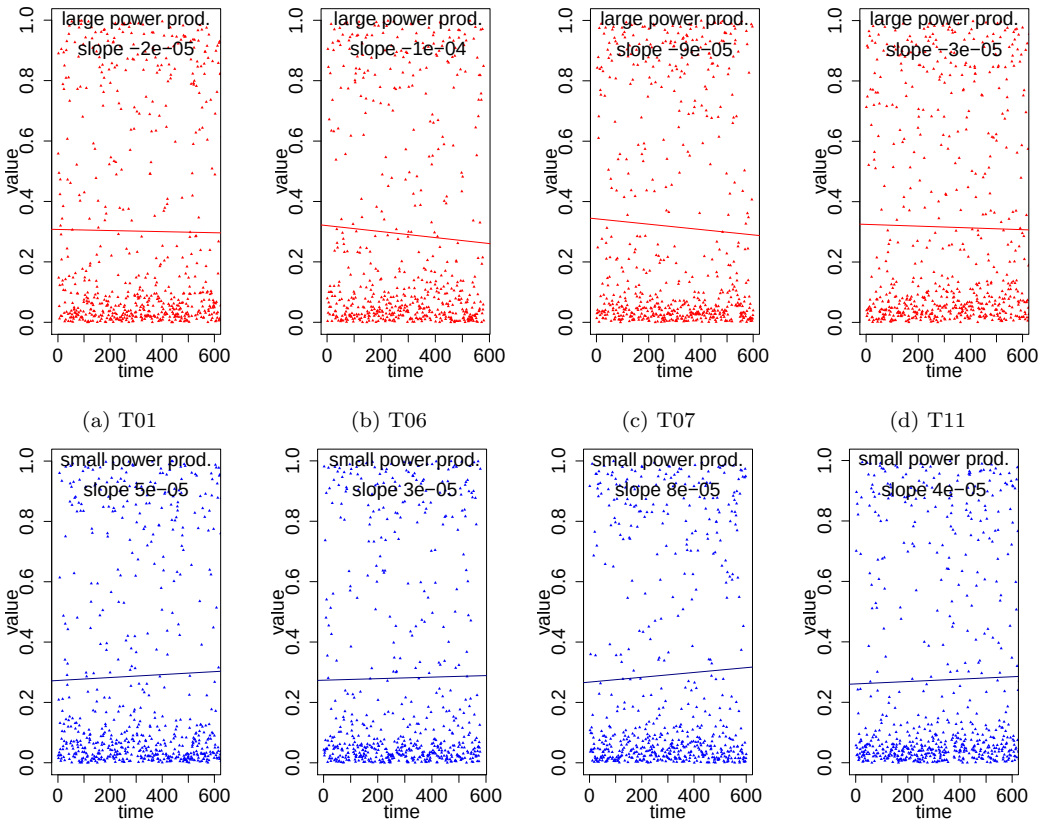
The plot concerns wind in the $[6.5, 7)$ range and temperature around 15 degrees Celsius for all wind turbines.

Figure 6 presents concept-based models created for T01, T06, and T07. For the sake of consistency in the simulations, we used the same experimental settings used to produce plots in Figure 3. Notice that we do not repeat the plot for T11 to avoid redundancy. The plots indicate a drift in power production. Notice that recent observations (yellow points) are shifted to the left, which means that recently there was less power produced than in the past for analogous environment conditions. The dispersion indicates the scale of this negative phenomenon. For example, the discrepancies are larger for T01 than T06.

4.3. Additional visual aid to evaluate aging

Finally, we would like to address the second method of turbine aging evaluation related to the visualization of the concepts in the two-dimensional space of value and change of value. We created three clusters for each window (we have Q windows) to represent the underlying raw values. The second proposed method uses extracted centroids and runs clustering for these centroids. Fuzzy c-means is executed separately three times for centroids representing *low*, *moderate*, and *high* concepts. Two new centroids are produced each time and describe the discrepancy in the concepts obtained in the first method. For each pair of new centroids, the further apart they are, the more differentiation we observed with time. We measured this differentiation with the Distance Index (DI) given in Equation (15).

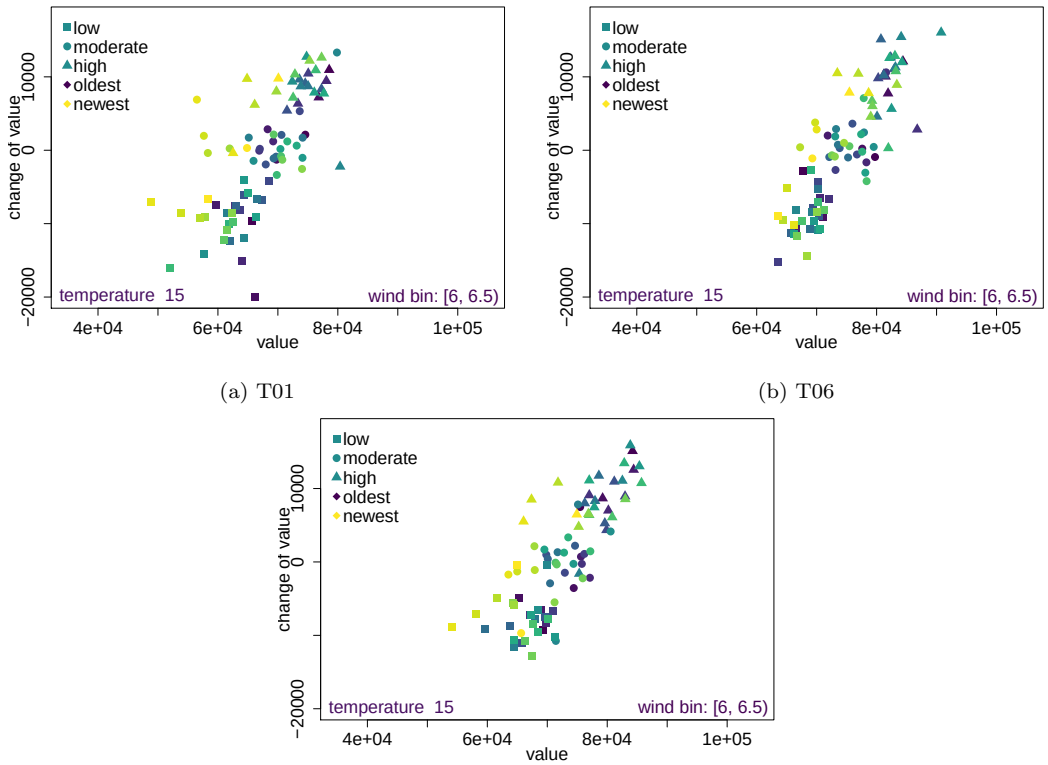
Figure 7 presents plots concerning the same wind speed and temperature conditions (wind speed $[6.5, 7)$ and temperature around $18°C$). We examine



(a) T01 (b) T06 (c) T07 (d) T11

(e) T01 (f) T06 (g) T07 (h) T11

Figure 5: Slopes computed for four wind turbines for the same weather conditions (wind speed $[6.5, 7)$ and temperature around $15°C$). Top row concerns memberships to the concept *high* power production. Here, negative values indicate aging. Bottom row concerns memberships to concept *low* power production, where a positive value shows aging. $Q = 30$



(a) T01

(b) T06

(c) T07

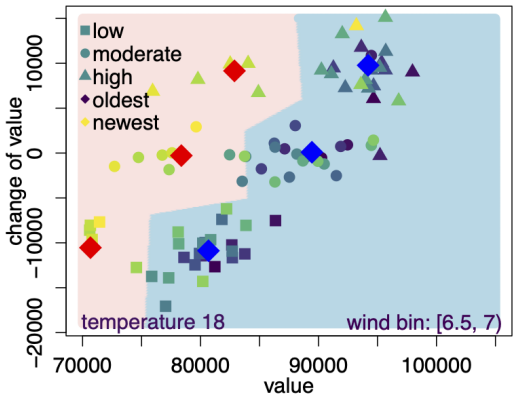
Figure 6: Concepts in the two-dimensional space of power value and change of power value in turbines (a) T01, (b) T06, and (c) T07. In each window, three concepts corresponding to *low* (squares), *moderate* (circles), and *high* (triangles) values were extracted. The most recent concepts are colored yellow. Plots concern wind speed $[6, 6.5)$ and temperature around $15°C$.

the performance of all four wind turbines in the EDP data set. The centroids used to compute the DI (Equation (15)) are marked with red and blue color diamonds. Blue is used to denote high power production ($\mathbf{v}_H^{(1)}, \mathbf{v}_H^{(2)}, \mathbf{v}_H^{(3)}$), and red is used to indicate low power production ($\mathbf{v}_L^{(1)}, \mathbf{v}_L^{(2)}, \mathbf{v}_L^{(3)}$). We colored the plot background so that the color informs which kind of concept is the closest for a given point in the plotted coordinate system. Pink indicates that the closest concept is red (low power), while light blue informs that the closest concept is blue (high power production). In addition, we report the DI values in the caption of each figure. These values were normalized as follows:

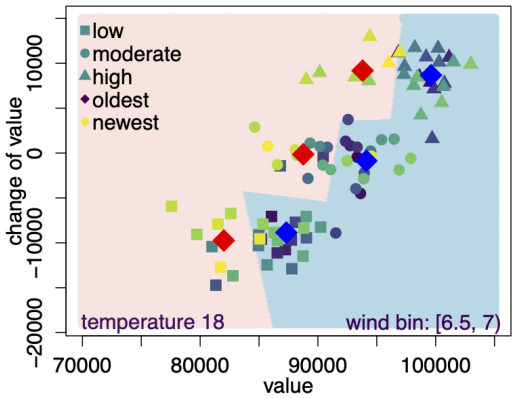
$$x_i = \frac{x_i - \min}{\max - \min} \tag{16}$$

where as *min* we used $40,000$ for the power value and $-20,000$ for the change of power for all wind turbines, both measured in Watt-hour. As *max* we used $100,000$ for the power value and $20,000$ for the change of power value for all wind turbines.

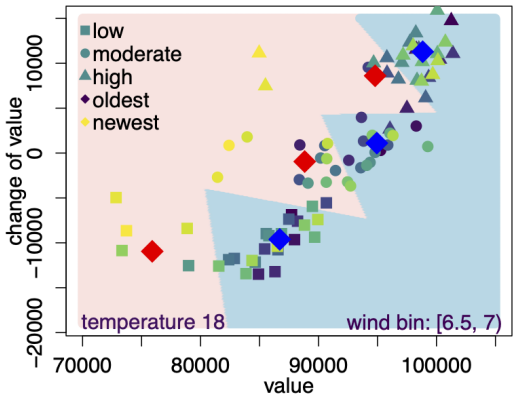
It is worth recalling that the DI is more informative when computed for more than one wind-temperature bin. Table 5 shows the normalized DI values calculated for a range of wind speed and temperature bins for all wind turbines in the data set.



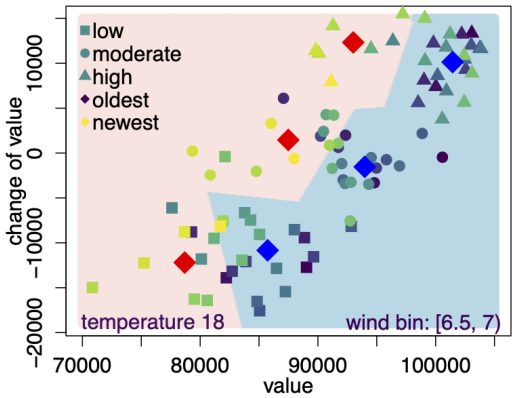
(a) T01, $DI = 0.59$



(b) T06, $DI = 0.29$



(c) T07, $DI = 0.39$



(d) T11, $DI = 0.40$

Figure 7: Secondary clustering applied to concepts extracted for turbines (a) T01, (b) T06, (c) T07, and (d) T11. Distance Index given by Equation (15) and normalized according to Equation (16) are displayed in sub-captions.

Table 5: Normalized DI values computed for four wind turbines for different weather conditions. The higher the index value, the worse the performance. $Q = 20$, $C = 3$.

Turbine ID	Temp. (°C)	Wind bins [5, 5.5)	[5.5, 6)	[6, 6.5)	[6.5, 7)	[7, 7.5)	**sum**
T01	15	0.52	0.38	0.51	0.32	0.74	**2.47**
	18	0.18	0.27	0.28	0.59	1.42	**2.73**
	22	0.15	0.28	0.41	0.42	0.47	**1.73**
	27	0.51	1.29	0.53	0.61	0.77	**3.71**
	sum	**1.36**	**2.22**	**1.73**	**1.94**	**3.40**	**10.65**
T06	15	0.29	0.34	0.62	0.95	1.55	**3.75**
	18	0.17	0.25	0.32	0.29	0.77	**1.80**
	22	1.07	0.28	0.40	0.39	0.44	**2.58**
	27	0.94	0.87	0.82	0.65	0.55	**3.83**
	sum	**2.47**	**1.74**	**2.16**	**2.28**	**3.31**	**11.96**
T07	15	0.22	0.29	0.37	0.34	1.45	**2.67**
	18	0.17	0.26	0.34	0.39	0.55	**1.71**
	22	0.99	0.69	0.28	1.25	0.82	**4.03**
	27	0.22	0.51	0.67	0.42	0.91	**2.73**
	sum	**1.60**	**1.75**	**1.66**	**2.40**	**3.73**	**11.14**
T11	15	0.19	0.28	0.41	0.36	0.41	**1.64**
	18	0.15	0.21	0.29	0.40	0.58	**1.62**
	22	0.50	0.31	0.30	0.34	0.33	**1.78**
	27	0.62	0.60	0.76	0.87	0.87	**3.73**
	sum	**1.46**	**1.40**	**1.76**	**1.97**	**2.19**	**8.78**

5. Conclusion

In the paper, we presented a methodology for visualizing and quantifying wind turbine performance degradation. Our approach uses the information of the following system variables: wind speed, atmospheric temperature, and generated power to assess the machine's health. The theoretical building

blocks of our proposal rely on fuzzy concepts labeled with symbolic terms. These fuzzy information granules represent and aggregate the data to capture power production shift patterns. The evaluation is performed separately for various operating conditions, as we recognize their vast influence on the machine performance.

Numerical simulations concerning four wind turbines in the EDP data set showed that the turbines with T01 and T11 deteriorated to a smaller extent than turbines with T06 and T07. Both approaches, regression-based and distance index-based, support this conclusion. The regression-based method provided more distinct differentiation in the turbines ranking, meaning that the discrepancies in turbines' scores were significant. However, we must note that this index is not expressed in natural units but membership degrees. In the future, we will be working on concept-based models for wind turbine failure description.

Acknowledgements

The project was partially funded by POB Research Centre for Artificial Intelligence and Robotics of Warsaw University of Technology within the Excellence Initiative Program - Research University (ID-UB).

Data Availability

All experiments were conducted using a dataset that can be downloaded from `https://opendata.edp.com/`.

References

[1] X. Jia, C. Jin, M. Buzza, Y. Di, D. Siegel, J. Lee, A deviation based assessment methodology for multiple machine health patterns classification and fault detection, Mechanical Systems and Signal Processing 99 (2018) 244–261.

[2] L. D. Avendano-Valencia, S. D. Fassois, Damage/fault diagnosis in an operating wind turbine under uncertainty via a vibration response Gaussian mixture random coefficient model based framework, Mechanical Systems and Signal Processing 91 (2017) 326–353.

[3] A. Reddy, V. Indragandhi, L. Ravi, V. Subramaniyaswamy, Detection of cracks and damage in wind turbine blades using artificial intelligence-based image analytics, Measurement 147 (2019) 106823.

[4] Y. Du, S. Zhou, X. Jing, Y. Peng, H. Wu, N. Kwok, Damage detection techniques for wind turbine blades: A review, Mechanical Systems and Signal Processing 141 (2020) 106445.

[5] M. Z. Jacobson, M. A. Delucchi, M. A. Cameron, B. V. Mathiesen, Matching demand with supply at low cost in 139 countries among 20 world regions with 100% intermittent wind, water, and sunlight (wws) for all purposes, Renewable Energy 123 (2018) 236–248.

[6] F. Zhang, Z. Wen, D. Liu, J. Jiao, H. Wan, B. Zeng, Calculation and analysis of wind turbine health monitoring indicators based on the relationships with SCADA data, Applied Sciences 10 (2020).

[7] C. Yang, J. Liu, Y. Zeng, G. Xie, Real-time condition monitoring and fault detection of components based on machine-learning reconstruction model, Renewable Energy 133 (2019) 433–441.

[8] K. F. Forbes, E. M. Zampelli, Accuracy of wind energy forecasts in great britain and prospects for improvement, Utilities Policy 67 (2020) 101111.

[9] X. Liu, K. Shi, H. Yu, Z. Zhu, Relative health index of wind turbines based on kernel density estimation, in: IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society, 2017, pp. 5957–5961.

[10] A. Kusiak, W. Li, The prediction and diagnosis of wind turbine faults, Renewable Energy 36 (2011) 16–23.

[11] J. Carroll, S. Koukoura, A. McDonald, A. Charalambous, S. Weiss, S. McArthur, Wind turbine gearbox failure and remaining useful life prediction using machine learning techniques, Wind Energy 22 (2019) 360–375.

[12] R. Wu, D. Zhang, Q. Yu, Y. Jiang, D. Arola, Health monitoring of wind turbine blades in operation using three-dimensional digital image correlation, Mechanical Systems and Signal Processing 130 (2019) 470–483.

[13] L. D. Avendano-Valencia, E. N. Chatzi, D. Tcherniak, Gaussian process models for mitigation of operational variability in the structural health

monitoring of wind turbines, Mechanical Systems and Signal Processing 142 (2020) 106686.

[14] D. Willis, C. Niezrecki, D. Kuchma, E. Hines, S. Arwade, R. Barthelmie, M. DiPaola, P. Drane, C. Hansen, M. Inalpolat, J. Mack, A. Myers, M. Rotea, Wind energy research: State-of-the-art and future research directions, Renewable Energy 125 (2018) 133–154.

[15] W. Yang, C. Liu, D. Jiang, An unsupervised spatiotemporal graphical modeling approach for wind turbine condition monitoring, Renewable Energy 127 (2018) 230–241.

[16] J. Zhan, R. Wang, L. Yi, Y. Wang, Z. Xie, Health assessment methods for wind turbines based on power prediction and Mahalanobis distance, International Journal of Pattern Recognition and Artificial Intelligence 33 (2019) 1951001.

[17] H. Ren, W. Liu, M. Shan, X. Wang, A new wind turbine health condition monitoring method based on VMD-MPE and feature-based transfer learning, Measurement 148 (2019) 106906.

[18] J. Zhang, N. Jiang, H. Li, N. Li, Online health assessment of wind turbine based on operational condition recognition, Transactions of the Institute of Measurement and Control 41 (2019) 2970–2981.

[19] S. Tewolde, R. Hoeffer, H. Haardt, Validated model based development of damage index for structural health monitoring of offshore wind turbine support structures, Procedia Engineering 199 (2017) 3242–3247.

[20] S. Koukoura, Failure and remaining useful life prediction of wind turbine gearboxes, in: Annual Conference of the PHM Society, volume 10, 2018, pp. 5957–5961.

[21] H. Ren, W. Liu, M. Shan, X. Wang, Z. Wang, A novel wind turbine health condition monitoring method based on composite variational mode entropy and weighted distribution adaptation, Renewable Energy 168 (2021) 972–980.

[22] J. Li, Q. Li, J. Zhu, Health condition assessment of wind turbine generators based on supervisory control and data acquisition data, IET Renewable Power Generation 13 (2019) 1343–1350.

[23] Z. Song, Z. Zhang, Y. Jiang, J. Zhu, Wind turbine health state monitoring based on a Bayesian data-driven approach, Renewable Energy 125 (2018) 172–181.

[24] K. Lopez de Calle, S. Ferreiro, C. Roldan-Paraponiaris, A. Ulazia, A context-aware oil debris-based health indicator for wind turbine gearbox condition monitoring, Energies 12 (2019).

[25] D. Tcherniak, Rotor anisotropy as a blade damage indicator for wind turbine structural health monitoring systems, Mechanical Systems and Signal Processing 74 (2016) 183–198.

[26] H. Zhao, H. Liu, W. Hu, X. Yan, Anomaly detection and fault analysis of wind turbine components based on deep learning network, Renewable Energy 127 (2018) 825–834.

[27] W. H. Lio, A. Li, F. Meng, Real-time rotor effective wind speed estimation using Gaussian process regression and Kalman filtering, Renewable Energy 169 (2021) 670–686. doi:https://doi.org/10.1016/j.renene.2021.01.040.

[28] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, A. Wu, An efficient k-means clustering algorithm: analysis and implementation, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 881–892.

[29] J. R. Dorrego, A. Rios, Q. Hernandez-Escobedo, R. Campos-Amezcua, R. Iracheta, O. Lastres, P. Lopez, A. Verde, L. Hechavarria, M.-A. Perea-Moreno, A.-J. Perea-Moreno, Theoretical and experimental analysis of aerodynamic noise in small wind turbines, Energies 14 (2021).

[30] J. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, International Journal of Cybernetics and Systems 3 (1973) 32–57.

[31] J. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Kluwer Academic Publishers, 1981.

[32] R. Hathaway, J. Bezdek, Fuzzy c-means clustering of incomplete data, IEEE Transactions on Systems, Man, and Cybernetics 31 (2001) 735–744.

[33] Y. Zhang, Z. Li, H. Zhang, Z. Yu, T. Lu, Fuzzy c-means clustering-based mating restriction for multiobjective optimization, International Journal of Machine Learning and Cybernetics 9 (2018) 1609–1621.

[34] B. A. Pimentel, R. M. C. R. de Souza, A generalized multivariate approach for possibilistic fuzzy c-means clustering, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 26 (2018) 893–916.