

Article

Clustering of Wind Speed Time Series as a Tool for Wind Farm Diagnosis

Ana Alexandra Martins ^{1,2} , Daniel C. Vaz ^{3,4}, Tiago A. N. Silva ^{1,3,4} , Margarida Cardoso ⁵
and Alda Carvalho ^{1,6,7,*} 

¹ Centro de Investigação em Modelação e Otimização de Sistemas Multifuncionais, ISEL/IPL, 1959-007 Lisboa, Portugal

² Centro de Investigação em Matemática e Aplicações, ISEL/IPL, 7000-671 Évora, Portugal

³ UNIDEMI, Department of Mechanical and Industrial Engineering, NOVA School of Science and Technology, Universidade NOVA de Lisboa, 1099-085 Lisbon, Portugal

⁴ Laboratório Associado de Sistemas Inteligentes, 4800-058 Guimarães, Portugal

⁵ Business Research Unit, ISCTE-IUL, University Institute of Lisbon, 1649-026 Lisboa, Portugal

⁶ Departamento de Ciências e Tecnologia, Universidade Aberta, 1250-100 Lisboa, Portugal

⁷ CEMAPRE/ISEG Research, University of Lisbon, 1269-001 Lisboa, Portugal

* Correspondence: alda.carvalho@uab.pt

Abstract: In several industrial fields, environmental and operational data are acquired with numerous purposes, potentially generating a huge quantity of data containing valuable information for management actions. This work proposes a methodology for clustering time series based on the K-medoids algorithm using a convex combination of different time series correlation metrics, the COMB distance. The multidimensional scaling procedure is used to enhance the visualization of the clustering results, and a matrix plot display is proposed as an efficient visualization tool to interpret the COMB distance components. This is a general-purpose methodology that is intended to ease time series interpretation; however, due to the relevance of the field, this study explores the clustering of time series judiciously collected from data of a wind farm located on a complex terrain. Using the COMB distance for wind speed time bands, clustering exposes operational similarities and dissimilarities among neighboring turbines which are influenced by the turbines' relative positions and terrain features and regarding the direction of oncoming wind. In a significant number of cases, clustering does not coincide with the natural geographic grouping of the turbines. A novel representation of the contributing distances—the COMB distance matrix plot—provides a quick way to compare pairs of time bands (turbines) regarding various features.

Keywords: time series; wind data; clustering; K-medoids; COMB distance; visual interpretation tools; wind farm diagnosis



Citation: Martins, A.A.; Vaz, D.C.; Silva, T.A.N.; Cardoso, M.; Carvalho, A. Clustering of Wind Speed Time Series as a Tool for Wind Farm Diagnosis. *Math. Comput. Appl.* **2024**, *29*, 35. <https://doi.org/10.3390/mca29030035>

Received: 15 February 2024

Revised: 27 April 2024

Accepted: 7 May 2024

Published: 9 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

1.1. Motivation

In the upcoming decades, wind power is expected to play an increasingly important role in the transition to a low-carbon energy system. Wind power is accessible and cost-competitive. Global wind electricity generation has been growing, reaching 2388 TWh in 2023 [1].

There has been ongoing research and development leading to efficiency and cost improvements in wind turbine technology. Most modern wind turbines are designed to have a useful life of around 20 to 25 years, with the actual lifetime depending on maintenance practices and economics. Many of Europe's onshore wind farms are reaching the end of their operational life: 14 GW of Europe's existing wind farms have already been operational for more than 20 years and 78 GW will have reached that milestone by 2030. Denmark, Spain, and Portugal have the oldest wind fleets when compared to the

other EU countries in relative terms. In 2022, the average age of their wind turbines was more than 12 years old [2]. This means that a period of significant wind farm repowering is approaching.

Wind farm repowering involves replacing old turbines with more powerful and efficient models that incorporate the latest technology. These newer models feature larger rotor diameters and hub heights. Consequently, what was previously generated by two, three, or even four turbines can now be produced by just one. As the relative distance between turbines, expressed in rotor diameters, must be maintained, the use of larger diameter rotors results in a distinct (and significantly reduced) number of turbines within the wind farm. Consequently, the arrangement of turbines over the land area must differ from the original configuration. This evolution is motivated by the potential to achieve a much higher power production from a given installation area, sometimes up to three times more.

Micro-siting is the process of identifying the best locations to place wind turbines within a wind farm to maximize energy production and reduce operational costs. This process entails identifying areas with higher wind speeds, which correspond to higher available power, and reduced turbulence in the wind, which is beneficial in terms of, e.g., power output stability and less wear and tear on turbine components.

In the context of wind turbine operation, a wind farm operator possesses data, including wind speed and direction recorded by anemometers on the wind turbines' nacelles. Turbulence intensity data may also be available depending on the sampling rate. Previous works [3,4] have proposed a data-driven approach to determine optimal locations for placing wind turbines within a wind farm. This approach involves leveraging historical data from a specific generation of wind turbines on the wind farm to enhance the micro-siting of the subsequent generation. This is particularly relevant as the next generation of turbines may differ in terms of the number of turbines and their nominal power, as mentioned earlier.

Valuable insights can be gained by analyzing real data to understand the distributions of wind speed (related to power output) and turbulence intensity (linked to maintenance costs) across the wind farm for all incoming wind directions. Toward that goal, the authors have previously introduced the concept of wind signature [4]. This graphical representation encompasses the time series of three key variables (wind speed, v ; turbulence intensity, TI ; and local wind direction, θ) for each wind turbine over a specified time frame, organized into so-called time bands. These time bands represent continuous data series meeting specific criteria and can possibly be gathered for any wind direction. Wind signatures for all turbines in a wind farm are then mapped, each corresponding to a time band and a relatively narrow sector of the wind direction reaching the farm. Sequences of these maps are visually compared to identify changes in flow patterns as wind interacts with the wind farm, providing insights into the impact of terrain features on the variables of interest.

Therefore, there is an opportunity for an approach, complementary to the wind-signature proposal, that can provide a comprehensive set of results depicting the aggregated or macroscopic distribution of wind characteristics throughout the wind farm. This approach aims to identify specific areas within the wind farm characterized by high or low values of wind speed. These areas would be determined by clustering turbines based on the similarity of wind time series recorded by their anemometers. The measure of similarity includes not only wind speed magnitude and fluctuation but also potential time lags between turbine signals. Grouping turbines in this manner facilitates the identification of zones within the wind farm with the highest or lowest wind potential.

Nevertheless, the challenge is intricate: each turbine on the wind farm provides a time series with multiple variables. Consequently, multiple pairs of time series must be compared, and it is necessary to quantify their similarity or dissimilarity using various metrics. There are numerous options for selecting these metrics. Clustering is a statistical method that can be used to analyze data and identify groups of entities or observations that share similar characteristics and that, in turn, differ from other groups.

1.2. Clustering Wind Farm Data

Clustering methods aim at discovering a set of homogeneous and well-separated groups (i.e., clusters) that constitute a partition (possibly a fuzzy or a probabilistic partition) of a sample of heterogeneous items. Many clustering approaches rely on the definition of a distance between items, based on their characteristics, e.g., the well-known Euclidean distance. Clustering methods for time series data have been used in many scientific fields [5].

In the context of wind farms, Al-Shammari and co-authors [6] compared a few clustering approaches, including K-medoids, for identifying “levels of wake effect patterns” on a wind farm. This study concluded that K-medoids performed better than the other two methods: fuzzy C-means and K-means. More recently, Adedeji and co-authors [7] used the K-means to cluster 44 wind turbines based on wind speed time series to assign a higher maintenance priority to turbine clusters with more wind availability and higher wear rate thereof.

1.3. Contribution and Content Overview

This work proposes a new methodology for clustering time bands using the K-medoids algorithm. Time bands represent uninterrupted time series of variables of interest for the study. The clustering process uses the COMB distance, which is a convex combination of four (normalized) distance measures: Euclidean distance, Pearson correlation based distance, periodogram-based distance, and autocorrelation-based distance. These four distances enable the comparison of pairs of bands by considering alternative features.

In this study, the proposed methodology is applied to a subset of wind data from a wind farm situated on a complex terrain, where various wind patterns have been identified [3]. When comparing wind speed time bands, differences in the Euclidean distance reveal changes in the intensity of the wind speed; differences in the Pearson based distance suggest changes in trends; differences in the periodogram-based distance are associated with turbulence intensity variations; and differences in the autocorrelation-based distance are linked to the time lag correlations between turbine signals. A graphical representation explicitly illustrating the contribution of each component distance provides invaluable insights.

This proposed methodology enables the study and assessment of the interactions of the wind approaching the turbines in a wind farm. Through clustering, we aim to reveal turbine groups within the wind farm, which can provide insights into the local orographic effects that impact turbine behavior. This understanding may inform future decisions regarding turbine placement.

The newly proposed matrix plot facilitates the identification of differences between wind farm turbines in a concise and informative manner. By associating these differences with the physical locations of turbine clusters within the wind farm, we anticipate gaining insights into how terrain features or turbine wakes impact the wind reaching the turbines. Consequently, understanding this influence can provide insights into their output or the mechanical loads to which they are exposed. The developed tools enable the diagnosis of operational features, supporting decision-making in wind farm management.

The rest of this paper is structured as follows: Section 2 presents the methodology, culminating in a summary illustrated in a flowchart (Figure 2). In Section 3, we delve into a case study, providing a detailed analysis of the results obtained from applying the methodology described earlier to a specific case. Section 4 provides final remarks and suggestions for future research directions.

2. Methodology

2.1. Harvesting Time Bands

The concept of time bands was introduced in a previous study [4] as an uninterrupted time series of variables of interest for the study—for example, data of wind variables and turbine operation parameters for all wind turbines under consideration—that comply

with certain criteria defined to carry out a specific filtering and analysis of the data. The criteria can be operationalized through a cumulative sequence of filters, e.g., considering a minimum value of wind speed at a given turbine (or at all turbines), a specific season of the year, ranges of wind direction, air temperature, turbulence intensity, etc.

Time bands were extracted from the entire time series of data collected at the wind farm, resulting in multiple bands of varying lengths. The length of a time series is crucial, as it should be sufficiently long to generate an ample number of time bands and possess enough duration to facilitate the intended analysis.

2.2. Clustering Time Bands

In this work, the K-medoids algorithm [8] was adopted to cluster time bands regarding wind speed time series data of a wind farm. In the K-medoids algorithm, the objective is to minimize (for all clusters) the distance between the time series belonging to a cluster and the cluster's medoid, a member of the data set that exhibits the smallest distance to all the other elements in a cluster. By considering a medoid, we overcome the need to determine a centroid based on averaging different time series, which can be problematic. Furthermore, K-medoids can handle multiple distance measures which is pertinent in our study. For determining the best number of clusters, we considered the Average Silhouette index [8]. This index varies between -1 (indicating a very poor model) and 1 (indicating an excellent model). An Average Silhouette greater than 0.7 indicates a strong structure has been found in the data, a value of this index between 0.5 and 0.7 indicates a reasonable structure, between 0.25 and 0.5 a weak structure, and an index less than 0.25 indicates that no substantial structure has been found [8].

2.3. A Combined Distance between Time Bands

Selecting a dissimilarity or a distance measure is a critical issue in clustering time series. We resorted to the COMB distance [9,10]. The COMB distance presents a good trade-off between performance and computation time when compared to alternative distances [10] and has been successfully used in other applications. In [9], the authors used the COMB distance to cluster time series of hourly prices of electricity. The COMB distance has also been used to identify daily load patterns for short-term load forecasting [11]. The COMB distance is a convex combination of four (normalized) distance measures: the Euclidean distance, a Pearson correlation based distance, a periodogram-based distance, and an autocorrelation-based distance.

Considering two time series x_t and y_t , ($t = 1, \dots, T$),

1. The Euclidean distance, d_{Eucl} , captures differences in scale,

$$d_{Eucl} = \sqrt{\sum_{t=1}^T (x_t - y_t)^2}. \quad (1)$$

2. The Pearson correlation based distance, $d_{pearson}$, was suggested by [12],

$$d_{pearson} = \sqrt{\frac{1 - r_{x_t, y_t}}{2}}, \quad (2)$$

with r_{x_t, y_t} representing the Pearson correlation between the time series x_t and y_t . It emphasizes the differences between trends over time.

3. The periodogram-based distance, d_{period} [13], is the Euclidean distance between the periodograms of the time series,

$$d_{period} = \sqrt{\sum_{j=1}^{\lfloor T/2 \rfloor} (P_x(w_j) - P_y(w_j))^2}, \quad (3)$$

where $P_x(w_j) = \left(\frac{1}{n}\right) \left| \sum_{t=1}^T x_t e^{-itw_j} \right|^2$ and $P_y(w_j) = \left(\frac{1}{n}\right) \left| \sum_{t=1}^T y_t e^{-itw_j} \right|^2$ are the periodograms' for x_t and y_t , respectively, at frequencies $w_j = 2\pi j/T, j = 1, 2, \dots, [T/2]$ in the range from 0 to π ($[T/2]$ being the largest integer less or equal to $T/2$). It captures the differences in the contributions of the various frequencies or cyclical components to the variability of the series.

4. The autocorrelation-based distance, $d_{Autocorr}$ [14], is the Euclidean distance between estimated autocorrelation functions,

$$d_{Autocorr} = \sqrt{\sum_{l=1}^L (r_l(x_t) - r_l(y_t))^2}, \tag{4}$$

where $r_l(x_t)$ and $r_l(y_t)$ represent the estimated autocorrelations of lag l of x_t and y_t , respectively. This measure stresses the differences regarding the dependence on past observations.

When comparing wind speed time bands, differences in the Euclidean distance indicate changes in the intensity of the wind speed. Variations in the way wind speed approaches turbines over time, such as changes in trends, can be captured by the Pearson based distance. Additionally, the behavior of the wind speed in the frequency domain (periodogram) and the time domain (autocorrelation) can be assessed using periodogram-based and autocorrelation-based distances, respectively. It is important to note that the frequency content of wind speed time bands is linked to its turbulence nature. Thus, higher values of the periodogram-based distance suggest higher turbulence intensity levels. Also, the autocorrelation-based distance addresses the time lag correlations between turbine signals. To encompass all these aspects, the COMB distance is a uniform convex combination of the four mentioned distances (min–max-normalized).

2.4. Classical Multidimensional Scaling (MDS) Procedure

As a complementary tool for the visualization of the turbine clusters obtained, a classical multidimensional scaling (MDS) was used [15] (we resorted to the method implemented in R described in [16]).

The goal of classical MDS is to represent, in a Euclidean space, a matrix of distances between points. The point coordinates are found in a way that the original distances are preserved. To be of practical use, only two-dimensional representation is considered. Thus, classical MDS represents a reduction in data dimensionality, and the proportion of variation in original distances explained by the representation using only two dimensions is used as a goodness-of-fit measure (GOF),

$$GOF = \frac{\sum_{i=1}^p \lambda_i}{\sum_{i=1}^{n-1} |\lambda_i|}, \tag{5}$$

where λ_i are the eigenvalues of $X'X$, and $X_{n \times p}$ represents the p coordinates of the n points to be represented. As in our case we used two dimensions, $p = 2$ in the numerator, and we considered the two largest eigenvalues. The GOF varies between 0 and 1 with a higher value corresponding to a better MDS map [16].

2.5. A Matrix Plot for COMB Distances

The COMB distance between pairs of time series embodies the diverse influence of each elementary distance—Euclidean, Pearson correlation based, periodogram-based distance, and autocorrelation-based—on the resulting value. A graphical representation, explicitly illustrating the contribution of each component distance, proves invaluable across various contexts. Following a previous work [17], we propose a COMB distance matrix plot. Each entry of the matrix plot (e.g., the one in Figure 1) contains, for a pair of time bands, information on the aggregated distance (via the diameter of each gray circle). Overlaid on the circles is a diagram featuring four orthogonal line segments, each representing the value of one of the four distances presented in Section 2.3, as follows: the red (north-

ward segment) corresponds to the Euclidean distance (d_{Eucl}), the blue (eastward segment) corresponds to the Pearson distance ($d_{Pearson}$), orange (southward segment) corresponds to the periodogram distance (d_{Period}), and green (westward segment) corresponds to the autocorrelation-based distance ($d_{Autocorr}$).

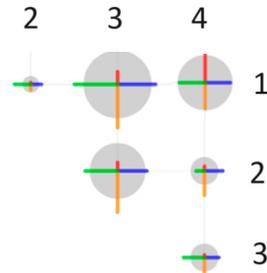


Figure 1. A COMB distance matrix plot. The diameters of the gray circles represent the COMB distance values. The length of the colored line segments refers to Euclidean (red), Pearson based (blue), periodogram-based (orange), and autocorrelation-based (green) distances. Each matrix element compares the two corresponding time bands. The depicted matrix compares four time series.

These colored line segments form a symmetrical cross “+” if the four distances are equal. Any lack of symmetry indicates that the contributions are unbalanced. The representation facilitates spotting such a situation and the identification of the component distance contributing the most to the (normalized) COMB distance.

The COMB distance matrix plot example in Figure 1 shows the combined distances between four time bands. One immediately noticeable aspect is the smaller circle for pair (1,2), indicating a stronger similarity. As the radius of these circles increases, greater dissimilarity is observed. The 4 distances can show identical contributions to the COMB distance—in the case of pair (1,4)—or show unbalanced contributions—in the case of pair (1,3). In the latter case, the Euclidean distance contributes less, while the periodogram distance contributes more to the final COMB distance. This visualization easily identifies uneven distance contributions that can provide valuable information, as each measurement has its interpretation. Although we developed this COMB matrix plot to study wind farms, the proposed graphical tool is not limited to this application.

2.6. Proposed Methodology in Brief

The proposed methodology may be summarized by its main procedures, as represented in the flowchart (Figure 2).

In the case of time series from wind turbines, the available data are usually acquired and managed by a SCADA (Supervisory Control And Data Acquisition) system. The SCADA data are raw data that need to be pre-processed, especially concerning data cleansing. After pre-processing, the wind data are then used to harvest time bands from the available time series of wind speed at angular sectors of interest of the wind direction. From the harvested time bands, uninterrupted time series of variables of interest are clustered using the K-Medoids algorithm based on the COMB distance. MDS is used to visualize the clustering results. A matrix plot is proposed as an efficient visualization tool to interpret the COMB distances between time bands.

The proposed methodology is here presented and applied for wind datasets, due to the relevance of this field of application, although this methodology can be applied for time series of numerous sources.

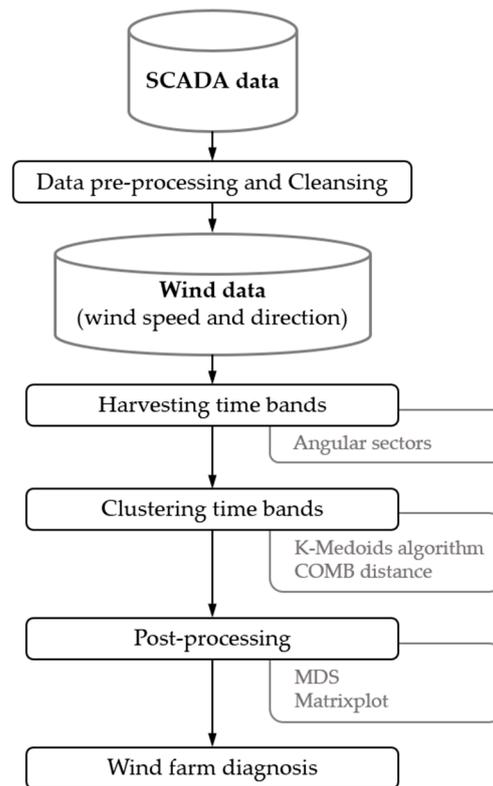


Figure 2. Flowchart with the main procedures of the proposed methodology.

3. Data Analysis and Results

3.1. Case Study

The wind farm taken as case study was in northern Portugal, on a plateau that rises 500 m above the surrounding terrain. Data available for this study originated from eight turbines (shown in Figure 3) amongst others that comprised the farm. There was a large separation of 1.45 km between turbines WT4 and WT5, leading to an intuitive clustering of the turbines into two geographical groups of four turbines each: G1, comprising turbines WT1 to WT4, and G2, comprising turbines WT5 to WT8. The average distance between neighboring turbines was approximately 280 m for group G1, and 290 m for group G2. Group G1 was at the edge of the plateau, with slopes due north, whereas group G2 was distant from steep slopes. There was no significant overgrown vegetation or construction obstacles on site. For group G1, high wind frequencies generally occurred in the south–southwest (SSW) sector and on a narrow sector centered at east–southeast (ESE), while very low wind frequencies occurred for the north–east or east directions, depending on the turbine [3].

The orography, and potentially vegetation growth, influence wind velocity magnitude and direction at specific turbine locations, depending on the incoming wind direction. These factors are crucial for power generation, as wind speed and fluctuations (turbulence) are the most significant features of the wind in this context. This wind farm was selected due to the identification of various wind patterns, which highlighted the effects of terrain features for different directions of incoming wind [3].

The data analyzed referred to the entire years of 2009 and 2010, with measurements taken every 10 min. Thus, the complete dataset contained 105,120 observations. The variables of interest in this study were wind speed, V , and wind direction, θ , which were measured by the anemometers installed on the wind turbines.

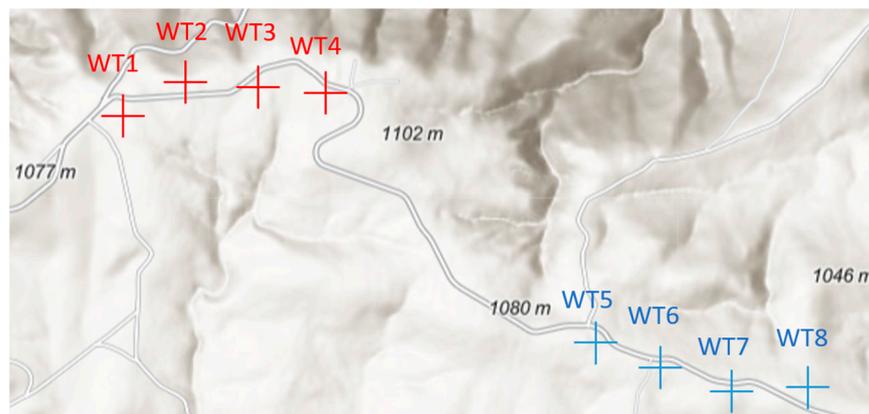


Figure 3. Wind farm: two groups, turbines WT1 to WT4 (red) and WT5 to WT8 (blue) are clearly identifiable according to the geographical distances between the wind turbines.

3.2. Data Pre-Processing

In diverse industrial applications, it is common to use SCADA systems to acquire and store data. Thus, SCADA systems manage the automation and synchronization of data collection and storage. In wind farms, data pertaining to operational and environmental variables may be collected with different acquisition rates. However, as a common practice for data related to wind turbines, like wind speed and direction, data are acquired at a somewhat fast rate (e.g., one data point per second) and then down-sampled to 10 min intervals between data points by averaging [4].

The data pre-processing or preparation step ensures compatibility among time series of different variables and wind turbines. In this work, after loading a dataset and defining the time frame, data vectors were stored in multidimensional arrays. Variables were initialized as arrays of “nan”, and for each variable, timestamp vectors of different turbines were compared. If they matched, variable vectors were recorded; otherwise, missing timestamp entries were identified, and existing data were recorded at registered timestamps, leaving unmatched entries as “nan”. The dataset was thus synchronized, containing compatible data entries, and it was saved in a single file. This preparation allowed for general-purpose treatment, including missing-data imputation techniques before cleansing and filtering. Another cleansing step was used to ensure that variables had values within their domain, such as out-of-range wind direction values being wrapped to the interval 0° to 360° . Note that the reported steps are general-purpose pre-processing steps and may be considered for different datasets. Regarding wind data, specific filters may be applied to the dataset to limit the analysis to physically meaningful observations, such as the restrictions related to the wind turbine operational range, namely, wind and turbine’s rotor speeds within given ranges. Other filters may be considered for specific applications such as the ones described to harvest time bands.

3.3. Harvesting Time Bands

Time bands for each turbine must be harvested to constitute the clustering base data. The time bands are uninterrupted time series of a variable, wind speed in the present study, referring to a specific wind direction sector. A sector is defined in terms of its mid-sector direction and angular aperture.

To provide reliable and informative data for clustering in the present study, they were selected according to the following criteria:

1. Wind speeds had to be above 3.5 m/s (63% of the initial dataset).
2. Wind direction as measured at a reference turbine had to be within the angular sector that had been predefined for the collection of a given set of time bands. Amongst the entire set, the reference turbine was the most upwind turbine in the mid-sector direction.

The first criterion left out periods of insufficient wind speed magnitude for power production. The second criterion helped spot more wind power availability within specific angular sectors. To implement this criterion, we considered 18 non-overlapping wind direction sectors of 20°. In Table 1, the 18 sectors are described. For each sector, the four largest time bands were kept for the analysis. They corresponded to completely distinct time periods. The length of the 72 selected time bands ranged from 3.3 h (length of 20 times 10 min) to around 34.2 h. The average size of the time bands was 51 (around 8.5 h). Table 1 also lists the reference turbine for each sector.

Table 1. The 18 wind direction sectors and the length of the 4 longest time bands obtained for each sector.

Sector ID	Sector		Clockwise Direction from North (°)	Most Upwind Turbine	Time-Band Length ¹			
	Start (°)	End (°)						
1	355	15	5	3	49	47	45	45
2	15	35	25	4	31	27	26	25
3	35	55	45	8	78	65	51	50
4	55	75	65	8	79	51	48	44
5	75	95	85	8	48	37	37	32
6	95	115	105	8	41	40	37	35
7	115	135	125	8	24	23	23	20
8	135	155	145	8	49	45	38	36
9	155	175	165	8	205	98	88	82
10	175	195	185	7	58	50	42	41
11	195	215	205	5	130	93	83	75
12	215	235	225	1	50	48	43	41
13	235	255	245	1	60	43	40	32
14	255	275	265	1	41	38	28	22
15	275	295	285	1	52	37	35	27
16	295	315	305	1	64	54	47	38
17	315	335	325	1	64	60	54	53
18	335	355	345	2	85	49	48	43

¹ Longest to shortest.

Figure 4 represents the time bands’ lengths in a circular plot. While in most sectors the length of time bands stayed around average, sectors 2 and 7 had shorter time bands and sectors 9 and 11 had longer time bands.

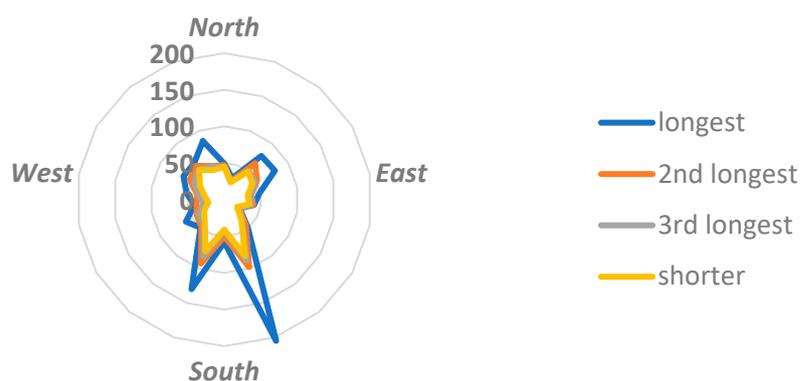


Figure 4. Time-band length by sector.

3.4. Clustering of Wind Speed Time Bands

In the previous section, 18 sectors were considered, and within each of these, the 4 longest were selected, resulting in a total of 72 time bands. Each time band consisted of eight time series of wind speed recorded at each of the eight turbines. We then conducted 72 cluster analyses (one for each time band) using the K-Medoids algorithm and the COMB distance.

The clustering analysis revealed two to six possible clusters (Table 2). Slightly more than half of the time bands exhibited two clusters, some of which corresponded to the geographic groups (highlighted in bold in Table 2). Only two of the analyzed time bands displayed a turbine distribution incompatible with the geographic division (highlighted with underline in Table 2).

Table 2. Number of clusters for each of the four longest time bands by sector.

Sector ID		Number of Clusters ¹			
1	5	3	2	2	
2	2	2	2	2	
3	2	3	2	2	
4	3	2	2	3	
5	2	4	4	3	
6	4	3	4	3	
7	6	2	2	2	
8	2	2	2	3	
9	4	3	5	3	
10	3	2	2	3	
11	2	2	2	2	
12	4	2	5	3	
13	2	2	5	3	
14	3	4	2	2	
15	4	4	2	5	
16	2	2	4	<u>4</u>	
17	<u>2</u>	3	2	<u>4</u>	
18	2	5	3	3	

¹ Longest to shortest time band.

In terms of clustering quality, the minimum Average Silhouette obtained was 0.4, indicating that a clustering structure was found for all conducted analyses. Approximately 67% of the clusters exhibited a reasonable clustering structure, with Average Silhouette values ranging between 0.5 and 0.7. Three results demonstrated a strong clustering structure, with an Average Silhouette above 0.7. Despite conducting a straightforward analysis of the relationship between clustering quality and the corresponding time-band lengths, no association could be identified. This suggests that harvesting shorter time bands does not necessarily lead to worse results in cluster analyses.

To get an idea of the distribution of range (max–min), Figure 5 illustrates the range of the number of clusters represented for each sector. Sectors 2 and 11 consistently exhibited two clusters across all time bands. Sector 7 displayed the most variability in the number of clusters.

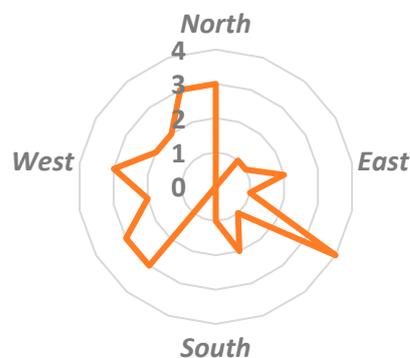


Figure 5. Range of number of clusters by wind direction sector.

Some clustering results obtained coincided with the two geography-based (“natural”) groups of turbines, as depicted in Figure 3 (highlighted in bold in Table 2). Clustering results

that clearly deviated from the “natural” clustering warranted a more thorough investigation, potentially leveraging information regarding the components of the COMB distance.

3.5. Classical MDS Representation

The MDS representation of clustering results offered a convenient method for quickly visualizing the COMB dissimilarities between turbines and the discovered clusters. In this representation, turbines were represented as points, and the COMB dissimilarities between them were depicted by the Euclidean distances on the map. Turbines that were closer together indicated more similar wind speed records according to the COMB metric.

In Figure 6, one can observe a solution comprising two clusters, corresponding to the two geographical groups depicted in Figure 3. This solution was derived from data of the second largest time band harvested for sector 11 (195–215°). The quality of that solution was assessed by an Average Silhouette index equal to 0.69, and the quality of the MDS representation was indicated by GOF = 0.82. The clustering results for sector 11 remained consistent across all time bands and aligned with the “natural” groups, as indicated in Table 2.

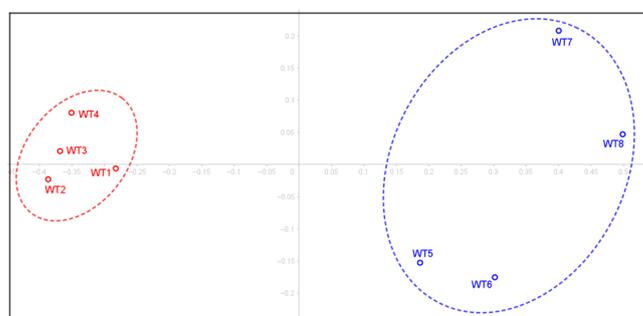


Figure 6. MDS representation with clustering solution obtained for sector 11, 2nd largest time band.

However, certain sectors, such as sector 9, yielded inconsistent clustering results across different time-band lengths. In Figure 7 (top), we illustrate the inconsistency of results derived from sector 9 data (second and third largest time bands). The solution depicted on the top panel has three clusters. It was derived from sector 9 and the second largest time band (of length 78). In that solution, the (“natural”) cluster including wind turbines WT5, WT6, WT7, and WT8 was subdivided into two clusters due to dissimilarities captured by the clustering procedure, indicating a distinct behavior in the wind approach to these groups. The quality of this clustering was measured by an Average Silhouette equal to 0.73 and the quality of the MDS map was measured by a GOF of 0.89. The MDS map presented on the bottom panel of Figure 7 refers to sector 9 data (third largest time band). For this MDS representation, the GOF was 0.88, and the quality of the five-cluster solution was measured by an Average Silhouette index equal to 0.48.

These clustering results suggested that for this wind direction sector (155–175), there were interactions between turbines and the terrain that were not consistent with the actual distribution of wind turbines on the wind farm. From the perspective of detecting an anomalous situation, the third largest time band of sector 9 was the most interesting to explore. The corresponding time series plots in Figure 8 (left), as well as the boxplot in Figure 8 (right) exhibit some dissimilarity in wind speed reaching the turbines. However, these tools did not reveal the clustering found in the MDS analysis (Figure 7, on the bottom). In fact, particular wind conditions across the wind farm may yield unexpected clustering results. However, for a better understanding, it is important to identify where the greatest dissimilarities are and what is causing them.

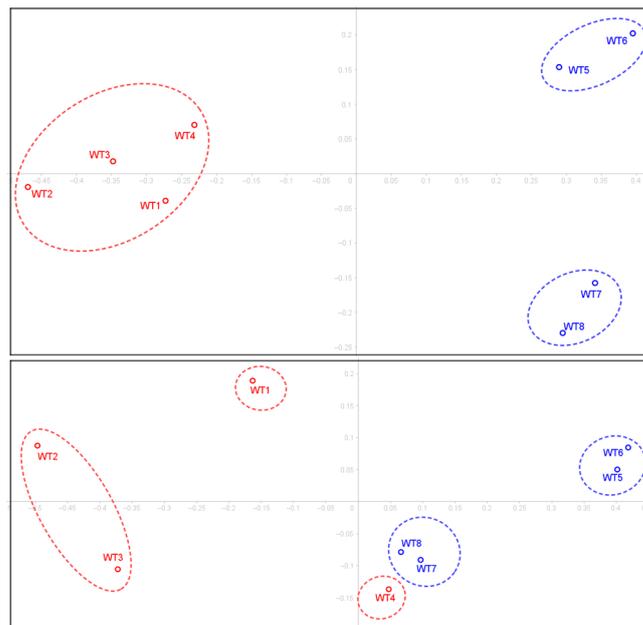


Figure 7. MDS representations with clustering solutions: analyses derived data from sector 9 (2nd largest time band at the **top** and 3rd largest time band at the **bottom**).

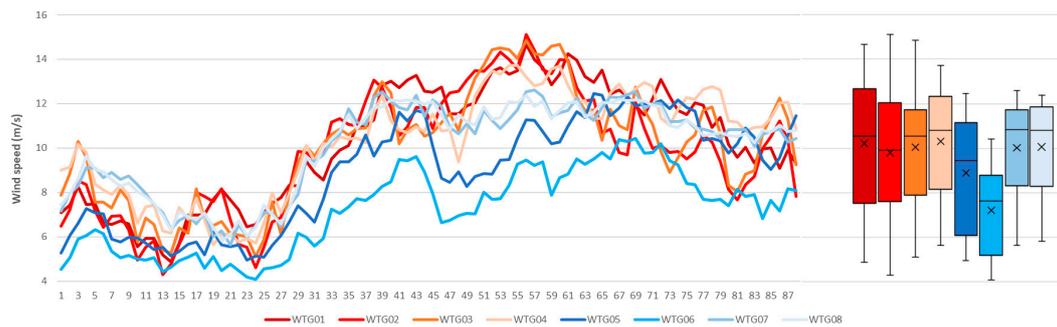


Figure 8. Time series plot (on the **left**) and boxplots (on the **right**) referring to time bands harvested in sector 9 (3rd largest time band).

In general, valuable insights can be gained from the clustering solutions by exploring which specific COMB distance might be contributing the most to the observed differences between wind turbines. To facilitate this assessment, we resorted to the proposed COMB distance matrix plot.

3.6. COMB Distance Matrix Plot between the Wind Farm Turbines

The purpose of this tool is to illustrate both the COMB distance and its components in a single plot. In Figure 9, the COMB distance matrix plot regarding the third largest time band of sector 9 is presented. We can verify that the pair (WT7, WT8) is very similar. Despite greater similarities within turbines in the same geographic group (smaller circles are mainly observed within the geographic groups), WT4 is also very similar to WT7 and WT8 (see also at the bottom of Figure 7 that these three turbines are very close). The largest circles are concentrated at the intersection of turbines in group G1 (WT1–WT4) with turbines in group 2 (WT5–WT8). Given the specific arrangement and turbine labeling of the wind farm, this manifests as a square pattern of large circles in the distance matrix. Large gray circle indicates a higher COMB distance, meaning that the turbines in the pair exhibit a distinct behavior in at least one of the aspects contributing to the COMB distance.

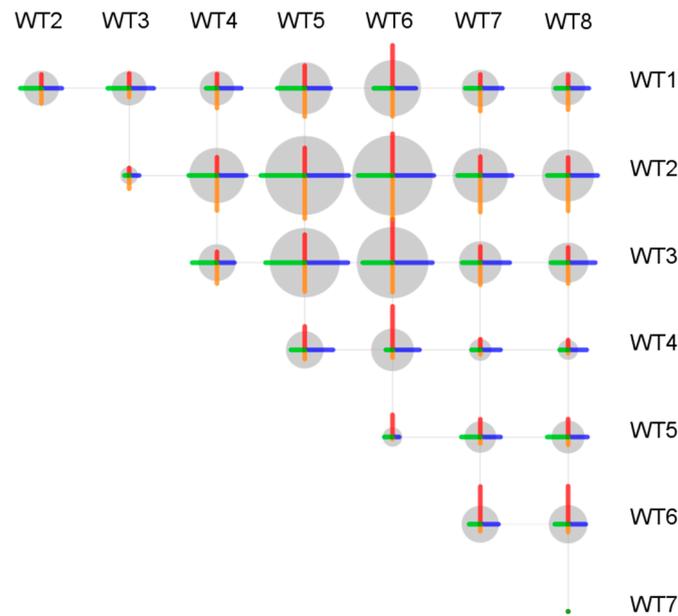


Figure 9. COMB distance matrix plot referring to sector 9, 3rd largest time band.

Also, the pair (WT1, WT5) displays a balanced distribution in the values of the four distances, evident in the four colored segments within the gray circle. However, this is not the case for the pair (WT4, WT6), for example, where the observed dissimilarity is primarily attributed to the Euclidean and Pearson distances.

As mentioned before, WT4 was very similar to WT7 and WT8, which was unexpected because WT4 belonged to a different geographical group than the other two turbines. In the matrix plot of Figure 9, the proximity between WT4 and the other two turbines are mostly due to the periodogram-based distance suggesting that despite belonging to a different geographical group, the turbine WT4 did not present great differences regarding turbulence intensity levels with the other two turbines.

Additionally, it can also be observed that the differences between WT4 and the other turbines were mostly due to the Pearson based distance, suggesting that the wind speed trends were in fact what set apart WT4. The time bands for the pair (WT4, WT7) are plotted in the bottom right panel of Figure 10.

To further explore the interpretation of the COMB distances, Figure 10 presents a more detailed analysis of some turbine pairs of time band 3 from sector 9. In this time band, the pair (WT7, WT8) presented the higher level of similarity (Figure 9) This similarity also appeared mirrored in the time series plot (Figure 10, top left). As the wind turbine speed varied between the turbine pairs, the distances values increased. However, it was not always easy to identify in the time series plot or boxplot what was causing this dissimilarity. For example, the pair (WT5, WT6) experienced the same type of wind but at different magnitudes. This fact is clearly visible in the time series plot as well as in the red segment in the distance matrix (Figure 10, top right).

There are certain pairs that exhibited dissimilarities caused by a reason other than wind speed magnitude. The other two examples in Figure 10 (bottom) show a dissimilarity mainly caused by correlation (blue segment, on the right) and by frequency (orange segment, on the left).

In the previous example, no dissimilarity caused by the autocorrelation structure was identified. To complete the analysis of the four distances, time band 3 of sector 12 was used to illustrate the ACF distance (Figure 11).

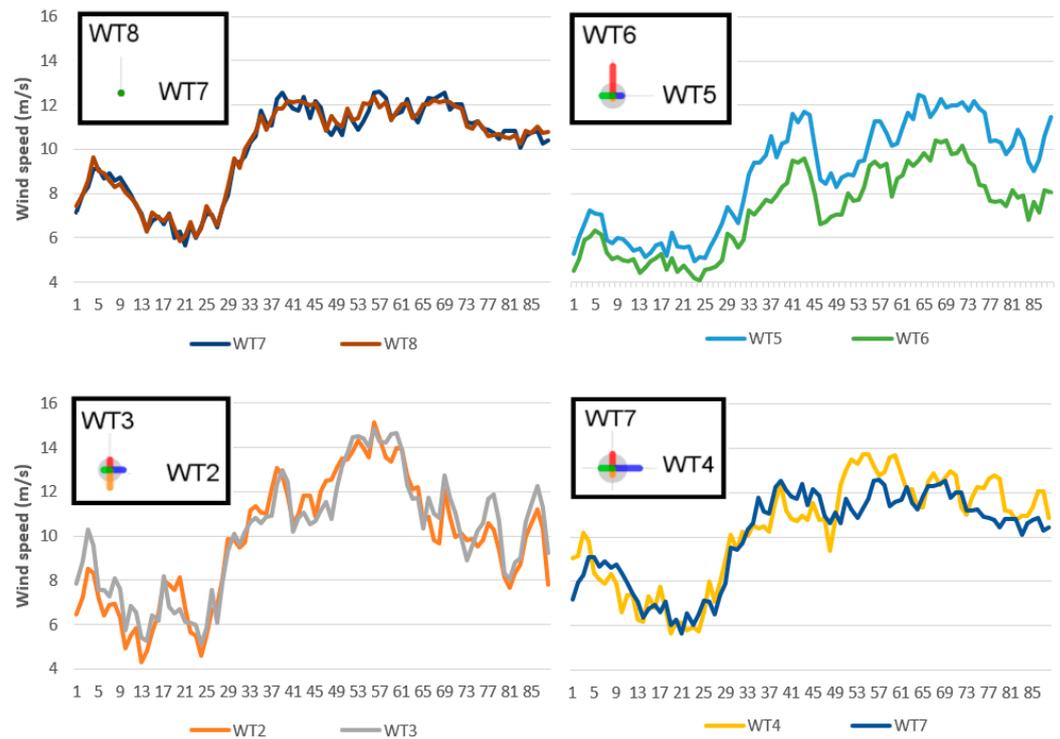


Figure 10. Time series plots and COMB distances between some pairs of wind turbines: analyses of data from sector 9, 3rd largest time band.

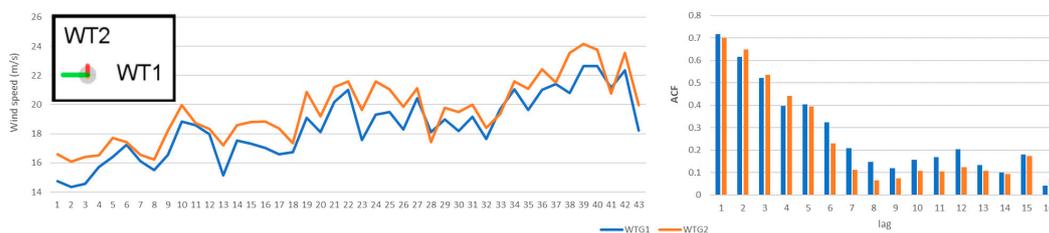


Figure 11. Sector 12, time band 3: time series plots and distances (left) and ACF (right).

Figures 10 and 11 show some selected time bands side to side to allow their visual comparison and their alignment with the COMB distance and its components. In Figure 10, it is evident that time bands for the pair (WT7, WT8) are not different, while the other pairs differ mostly in a specific distance, capturing specific differences of the time series and highlighting their contribution to the COMB distance. Time bands for the pair (WT1, WT2) in Figure 11 differ most significantly in their ACF (autocorrelation-based distance).

4. Final Remarks and Future Work

The quantity of data acquired by SCADA systems in wind farms is huge. These data contain valuable information for a successful wind farm operation and management. However, dealing with a huge quantity of data is usually a challenge that may lead to a waste of information due to a lack of data interpretation capabilities.

To bridge this gap, a methodology was proposed for clustering time bands based on the K-medoids algorithm, employing a convex combination of various time series distances measures known as the COMB distance. This approach incorporated information that enabled the assessment of similarities and dissimilarities in pairs of time bands, capturing differences in scale, trend, variability, and autocorrelation. Furthermore, multidimensional scaling (MDS) enhanced the visualization of clustering results, and a matrix plot display was introduced as an efficient tool for interpreting the COMB distance results for each pair of wind turbines on the wind farm at a glance.

Indeed, the proposed methodology was introduced and implemented specifically for wind datasets, given the significance of this field. However, its capability for interpreting time series can be extended to various sources. It is important to highlight that additional variables, such as turbulence index or air temperature, could be incorporated into the time bands analysis. Depending on the study's objectives, alternative dimensions could replace or complement wind direction in both criteria.

In this work, clustering based on K-medoids was applied to a two-year wind dataset, corresponding to SCADA data from two groups of four wind turbines on a wind farm on a complex terrain. The harvested time bands were the input for the clustering. As there were two geographically distinct groups of four wind turbines on the wind farm, it was expected that the time bands' clustering would correspond to these two groups. However, the number of obtained clusters varied from two to six, and even when two clusters were identified, they often were composed of groups of wind turbines that differed from the ones related to their geographical arrangement. The analysis of the wind turbines in each cluster allowed for the assessment of the wind reaching the farm from specific wind direction sectors, which contributed to the diagnosis of the wind flow on the wind farm and its interaction with the terrain. The COMB distance computed for time bands harvested for specific wind direction sectors exposed that the wind reaching the farm from a specific sector could result in dissimilar wind records at the various turbines. This was related to the interactions between turbines and the terrain that were not consistent with the actual distribution of wind turbines in the wind farm. In the context of wind data, the clustering results may be used to perform anomaly detection via the analysis of dissimilarities between pairs of time bands related to different wind turbines supporting the decision-making process regarding different aspects of the wind farm management over its lifecycle. Hence, by interpreting the COMB distance, one may track trends in specific distances relating them to changes in time series, e.g., a trend on the periodogram-based distance is related to changes in variability, empowering the analysis of information that is valuable to the wind farm management, regarding maintenance management, and giving information on the actual wind flow on the wind farm.

The presented results are promising, suggesting the potential of the application of the proposed methodology to treat time series from further variables of interest. For future developments, one anticipates that the study of turbulent intensity time series and its correlation with results in this study will allow for the assessment of changes in terrain features and for the correlation of the wind reaching each wind turbine with maintenance performance indicators.

Author Contributions: Conceptualization, A.C., A.A.M., T.A.N.S. and D.C.V.; computational modeling and statistical analyses, A.C., A.A.M. and M.C.; funding acquisition, A.C. All authors contributed equally to the remaining work. All authors have read and agreed to the published version of the manuscript.

Funding: This work was developed and financially supported under the framework of project IPL/2022/V5_FGM_ISEL. D.C.V. and T.A.N.S. acknowledge Fundação para a Ciência e a Tecnologia (FCT-MCTES) for its financial support via projects UIDB/00667/2020 and UIDP/00667/2020 (UNIDEMI). Margarida Cardoso is a BRU-IUL member and had the financial support of Fundação para a Ciência e a Tecnologia (FCT-MCTES) via grant UIDB/00315/2020 (DOI: 10.54499/UIDB/00315/2020). A.C. was partially supported by the Project CEMAPRE/REM—UIDB/05069/2020—financed by FCT/MCTES through national funds.

Data Availability Statement: Data is unavailable due to privacy.

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

References

1. IEA. *Wind Electricity*; IEA: Paris, France, 2022. Available online: <https://www.iea.org/reports/wind-electricity> (accessed on 15 February 2024).
2. Wind Europe. Repowering Europe's Wind Farms Is a Win-Win-Win. 2022. Available online: <https://windeurope.org/newsroom/press-releases/repowering-europes-wind-farms-is-a-win-win-win/> (accessed on 15 February 2024).

3. Casaca, C.; Vaz, D.; Silva, T.A.N.; Carvalho, A. An analysis of wind farm data to evidence local wind pattern switches near a plateau. In Proceedings of the 4th International Conference on Numerical and Symbolic Computation: Developments and Applications, Porto, Portugal, 11–12 April 2019; ISBN 978-989-99410-5-2.
4. Carvalho, A.; Vaz, D.C.; Silva, T.A.N.; Casaca, C. A Methodology to Reveal Terrain Effects from Wind Farm SCADA Data Using a Wind Signature Concept. In *Recent Developments in Statistics and Data Science*; Bispo, R., Henriques-Rodrigues, L., Alpizar-Jara, R., de Carvalho, M., Eds.; SPE 2021, Springer Proceedings in Mathematics & Statistics; Springer: Cham, Switzerland, 2022; Volume 398.
5. Aghabozorgi, S.; Seyed Shirshorshidi, A.; Wah, T.Y. Time-series clustering. A decade review. *Inf. Syst.* **2015**, *53*, 16–38. [[CrossRef](#)]
6. Al-Shammari, E.; Shamshirband, S.; Petković, D.; Zalnezhad, E.; Lip-Yee, P.; Suraya-Taher, R.; Čojbašić, Ž. Comparative study of clustering methods for wake effect analysis in wind farm. *Energy* **2016**, *95*, 573–579. [[CrossRef](#)]
7. Adedeji, P.; Olatunji, O.; Madushele, N.; Akinlabi, S.; Adeyemo, J. Cluster-based wind turbine maintenance prioritization for a utility-scale wind farm. *Procedia Comput. Sci.* **2022**, *200*, 1726–1735. [[CrossRef](#)]
8. Kaufman, L.; Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2009. [[CrossRef](#)]
9. Cardoso, M.G.M.S.; Martins, A.; Lagarto, J. Combining various dissimilarity measures for clustering electricity market prices. In *Estatística: Desafios Transversais às Ciências dos Dados—Atas do XXIV Congresso da Sociedade Portuguesa de Estatística*; Milheiro, P., Pacheco, A., de Sousa, B., Alves, I.F., Pereira, I., Polidoro, M.J., Ramos, S., Eds.; SPE: Richardson, TX, USA, 2021; pp. 197–212.
10. Cardoso, M.G.M.S.; Martins, A.A. The performance of a combined distance between time series. In *Recent Developments in Statistics and Data Science*; Bispo, R., Henriques-Rodrigues, L., Alpizar-Jara, R., de Carvalho, M., Eds.; SPE 2021, Springer Proceedings in Mathematics & Statistics; Springer: Cham, Switzerland, 2021; Volume 398. [[CrossRef](#)]
11. Martins, A.A.; Lagarto, J.; Canacsinh, H.; Reis, F.; Cardoso, M.G.M.S. Short-term load forecasting using time series clustering. *Optim. Eng.* **2022**, *23*, 2293–2314. [[CrossRef](#)]
12. Rodrigues, P.P.; Gama, J.; Pedroso, J. Hierarchical Clustering of Time-Series Data Streams. *IEEE Trans. Knowl. Data Eng.* **2008**, *20*, 615–627. [[CrossRef](#)]
13. Caiado, J.; Crato, N.; Peña, D. A periodogram-based metric for time series classification. *Comput. Stat. Data Anal.* **2006**, *50*, 2668–2684. [[CrossRef](#)]
14. Montero, P.; Vilar, J. TSclust: An R package for time series clustering. *J. Stat. Softw.* **2014**, *62*, 1–43. [[CrossRef](#)]
15. Torgerson, W.S. Multidimensional scaling: Theory and method. *Psychometrika* **1952**, *17*, 401–419. [[CrossRef](#)]
16. Cox, T.; Cox, M. Multidimensional Scaling. In *Handbook of Data Visualization*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2000; pp. 315–347.
17. Martins, A.A.; Vaz Daniel, C.; Silva Tiago, A.N.; Cardoso, M.G.M.S.; Carvalho, A. An application of time series clustering using a combined distance. In Proceedings of the 6th International Conference on Numerical and Symbolic Computation Developments and Applications, Évora, Portugal, 30–31 March 2023; ISBN 978-989-99410-7-6.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.