Spatio-temporal Analysis of Wind Resource in the Iberian Peninsula with Data-coupled Clustering

Mihaela I. Chidean[a], Antonio J. Caamaño[a,*], Julio Ramiro-Bargueño[a],
Carlos Casanova-Mateo[b], Sancho Salcedo-Sanz[c]

[a]*Dept. of Signal Theory and Communications, Universidad Rey Juan Carlos, Madrid, Spain.*
[b]*Dept. of Civil Engineering: Construction, Infrastructures and Transports,*
Universidad Politécnica de Madrid, Madrid, Spain.
[c]*Dept. of Signal Processing and Communications, Universidad de Alcalá, Madrid, Spain.*

Abstract

In this paper a spatio-temporal analysis of wind power resource in the Iberian Peninsula is presented. The study uses the Second-Order Data-Coupled Clustering (SODCC) algorithm over reanalysis data in the for the period 1979 – 2014. Several characteristics of the method are detailed, such as the data-coupled clustering approach of SODCC, that ensures the non-singularity of the signal subspace within each cluster. The performance of the proposed approach and specific results obtained have been discussed in a case study in the Iberian Peninsula. In these results it is possible to identify different spatio-temporal patterns of the wind data statistics depending on the initialization year. Moreover, this work also shows that there is a close relationship between these spatio-temporal patterns with the wind energy production of the area under study, so the proposed analysis can be extended to wind farms efficiency production at the time scales considered.

Keywords: Wind resource, Wind energy production, Spatio-temporal analysis, SODCC clustering

1. Introduction

In the last years, the growing of global wind power generation has been spectacular, with total investments over 100 billion dollars, and over 432GW of total installed power by the end of 2015 [1]. Wind power, together with alternative renewable energy sources, offers important benefits compared to fossil fuels, including a drastic reduction of emissions, environmental impact and climate change. On the other hand, wind intermittent nature is a major issue to achieve an effective integration of wind energy within existing power grids [2]. In close connection with its intermittent nature, wind resource assessment is a major issue for wind energy development, in different aspects such as wind farms prospection, analysis of current production in existing facilities, or as a tool for future projection of wind energy

[*] Corresponding author: `antonio.caamano@urjc.es`

production. This production heavily depends on weather and atmospheric conditions, which introduces an important spatio-temporal variability in the available wind resource [3].

The research on spatio-temporal analysis and structure of wind variability has been intense in the last years. In [4] the temporal changes in wind speed (10m) and direction has been analyzed in the North Sea. Different runs of several regional climate models are used to obtain conclusions on short-term correlations and also long-period fluctuations. In [5] the spatio-temporal variation of offshore wind resources is studied, based on remote sensing data from instruments onboard satellites. Wind field data from different satellites in the period 2000-2008 are used to carry out this characterization and study of the offshore wind resource potential in China. In [6] temporal and spatial variability of wind resource in the USA is analyzed. That paper uses the climate forecast system reanalysis to carry out this spatio-temporal analysis of wind speed data from 1979 to 2011. Different variation patterns of wind in the USA associated with climate indices such as the North Atlantic Oscillation (NAO) or El Niño South Oscillation (ENSO) are described. In [7] a study of the wind potential in Malaysia (at nationwide scale) is carried out. More specifically, that paper analyzes spatio-temporal variations of wind potential and spatial wind power density in Malaysia, by means of applying different models to spatially reconstruct the wind speed, and analyze its evolution within the time. In [8] an analysis on how the optimal siting of wind farms based on a pre-assessment of the spatio-temporal variability of wind resources could reduce fluctuations in the delivered output, improving this way the quality of the supply in the future. An analysis in Spain reveals the importance of taking into account this spatio-temporal evolution of wind resource in the location of wind farms. In [9] real-world historical wind data are used to analyze spatio-temporal correlation among several wind farms, and how this correlation has an impact on a competitive pool market, when inter-temporal constraints of dispatchable generating units are considered in the market model. In [10] an analysis of long-term wind speed evolution in Australia is carried out. In this case, a large number of wind observations and also reanalysis products are analyzed to compute such long-term wind speed trends. Spatio-temporal properties of wind speed have been also exploited to obtain a better wind speed prediction, such as in [11], where the Wavelet Transform is used for the decomposition of the wind speed data into more stationary components, and then a spatio-temporal model on each subseries is applied, for incorporating both temporal and spatial information.

Different clustering techniques have also been applied to wind resource analysis and prediction. In [12] use the furthest neighbour hierarchical clustering method to group the temporal instants of wind speed data in a unique meteorological station in Mexico. In [13] the k-means clustering technique is used to perform short term-prediction of the wind power at low wind speed using 10s data. From the multiple simulated scenarios, with different parameter combinations, their results show that for the customized cluster parameters the prediction accuracy can be improved. In [14] the k-means clustering technique is also used to evaluate and to forecast of the energy deliverable by different renewable sources. The main difference is that both wind speed and solar irradiance are considered as input data in association to the energy capability of a wind generator and a photovoltaic plant. In [15] an assessment of the renewable energy potential in Romania using the k-means algorithm is carried out. In this case, the clustering approach is applied to different variables related to the renewable energy potential in the country, such as installed capacity, level voltage, type of renewable technology and geographical location of the resource. Representative pattern areas of renewable energy sources are then obtained with the clustering method, which can be used to improve planning and development of the electric networks in Romania. [16] investigates the wind speed forecasting using a spectral clustering method, decomposing the wind speed data by means of the Wavelet Transform. With this method, the prediction accuracy for simultaneously time instants is increased. In [17] different clustering techniques are applied to obtain patterns of the wake effects in wind farms. Specifically, FCM (Fuzzy c-means), K-mean, and K-medoids were used as the clustering algorithms, and a Principal Component Analysis technique was previously applied as a pre-processing step to the wake data. Ten different wake effect clusters affecting the wind farm production were observed according to results obtained in that paper. [18] proposes to apply a cluster analysis to the numerical weather prediction information, in such a way that similar days in terms of wind power receives a similar prediction treatment. The prediction results show that correct cluster analysis method is a useful tool in day-ahead wind power prediction. The main drawback of these papers is that their analysis is limited to either the temporal or spatial scales.

This paper deals with a problem of spatio-temporal evaluation of wind speed, in connection with wind energy production and resource. Specifically, a novel data-coupled clustering technique is used to carry out this spatio-temporal analysis, over the ERA-Interim [19] reanalysis data in the Iberian Peninsula. The proposed method is based on the Second-Order Data-Coupled Clustering (SODCC) algorithm, that clusters the nodes of a given network using the statistics the measured data. In this work, the network elements are the output nodes of the ERA-Interim reanalysis in the area of interest. The application of the proposed method to wind speed reanalysis data in the Iberian Peninsula is justified, since Spain and Portugal are two of the major wind energy producers/consumers in the World [1]. The obtained results show two different types of behaviour of the wind resource in the area of interest, leading to different types of wind energy production patterns.

The rest of this paper is organized as follows: Section 2 describes the spatio-temporal analysis method based on the SODCC algorithm and its implementation; Section 3 describes the experiments made using wind data, presents our results and validate the analysis method; Section 4 discusses and interprets the obtained results. Finally, Section 5 concludes the article.

2. Methods

As mentioned, the main goal of this work is the presentation of an advanced signal processing method for wind speed data analysis, with promising applications for the study of wind farms efficiency. The different steps included in this data analysis method can be summarized as follows:

- Model the environment under study according to a given system model.

- Cluster the network using the SODCC algorithm with multiple temporal initializations.
- Calculate the cluster size probabilities, in order to study the temporal evolution of the statistics of the data. At this point, it is possible to identify distinctive patterns between different SODCC initializations by using the Kullback-Leibler divergence (KLD) metric.
- Calculate the node to cluster size probabilities, in order to spatial extent of the statistics of the data.

The following sub-sections detail this data analysis method and the required algorithms and mathematical tools.

2.1. System model

Consider a set of points located in a geographical region and a given physical variable measured over the time for each location. Each point is modeled as a node forming part of a network, modeled by the graph $G = (V, E)$, with $|V| = N$ vertexes/nodes and $|E| = C$ edges/connections.

Each of the N node measures a new data point every F_s time instants obtaining a total of M measurements per node. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_M]$, with $\mathbf{x}_m \in \mathbb{R}^N$, be the dataset measured by the entire network. The covariance matrix of $\mathbf{X}$ can be estimated as $\mathbf{\Sigma} = \mathbf{X}\mathbf{X}^\top/M$, where $(\cdot)^\top$ indicates the transpose operation.

2.2. SODCC: a data-coupled clustering algorithm

The SODCC algorithm is a data-coupled clustering algorithm that organizes the nodes of a network in terms of the second-order statistics of the measured data. SODCC was proposed for Wireless Sensor Network (WSN) clustering, but it is also suitable for the considered system model and for climate data analysis [20, 21].

In short, SODCC operates as follows: each non-final (or non-stable) cluster fusions to another cluster until a stopping criterion is reached. SODCC decides based on the relation between the dimension of signal subspace of the data ($\hat{d}$) and the cluster size (N_i). With this approach, it ensures the non-singularity of the signal subspace of the data measured by each final cluster, meaning that $\hat{d} < N_i$ and $\mathbf{\Sigma}$ is well-posed. In other words, SODCC determines that a cluster is final if it is possible to solve the principal components that explain most of the variance in the data (e.g. 90%).

The operation of the SODCC algorithm is divided in two stages: 1) random initialization of the first Cluster Heads (CHs) and formation of the first clusters, and 2) fusion between existing clusters set by the decision criterion. In this work, the dimension of the signal subspace $\hat{d}$ for the data in each cluster is calculated by means of the Fast Subspace Decomposition (FSD) [22].

Previous work have shown the relation between the decision criterion of SODCC and the final cluster sizes [20]. This is also illustrated in Figure 1. Given the final cluster configuration, it is expected that small clusters will appear in areas where the cross-correlation of the data among those nodes is high, as only data from few nodes are needed to extract the

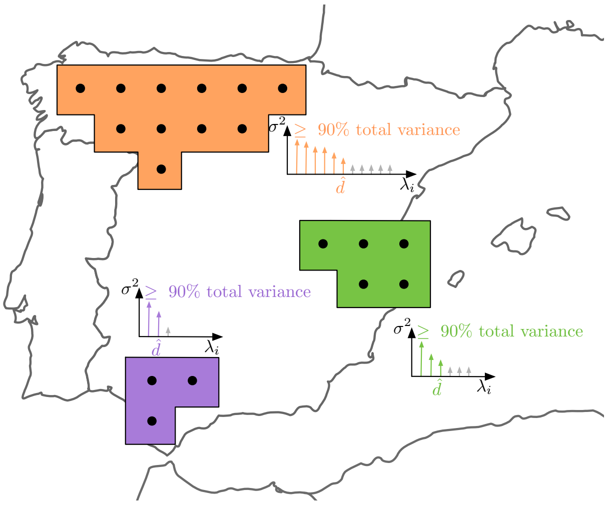


Figure 1: The cluster configuration given by SODCC can be understood in terms of the explained variance. The data variance from nodes in small clusters (purple) can be explained with almost all of the (few) present principal components. As clusters enlarge, more principal components are needed to explain the variance of the measurements (green and orange). In this figure, the graphs schematize the eigenvalues λ_i sorted in terms of their power σ^2 for each of the three clusters represented; the $\hat{d}$ eigenvalues that explain the 90% of the variance in each case are highlighted.

largest principal components that explain their variance. On the other hand, large clusters will appear in areas where the cross-correlation of the data among nodes is low. Therefore, SODCC ensures that closely correlated data series are clustered together and that the cluster configuration is linked to the measured field. A more detailed explanation of this clustering algorithm can be found in [20].

2.3. Wind speed data analysis with SODCC

This signal processing method to analyze climatological data using the SODCC algorithm starts considering a network modeled by a given graph G, starting to measure the physical variable in the time origin t_0. By applying SODCC to this setting repeatedly and independently, multiple final clustering configurations are obtained; and by performing statistical analysis, the underlying spatial structure of the physical variable is revealed. The conclusions from this analysis are obtained by studying the spatial extent and localization of the clusters and their most probable sizes. For example, areas where clusters of a given size appear with a high probability reveal the **spatial extent** (or scale) of the data correlation showing where the statistics of the measured data are similar.

The temporal structure of the measured data can be analyzed considering multiple time instants t_I where SODCC starts to cluster the network, i.e.

$$t_I \in \{t_j | j = 0, 1, 2, \dots\} \tag{1}$$

The definition of t_I depends on the final goal of each study and on the considered sampling frequency F_s. For example, for a monthly-scale study, the first time stamp in each month is a suitable element of t_I, and F_s should be high enough to allow SODCC to operate using only the data points of interest (e.g. measured during that month). For this setting, differences in the results obtained for the different t_j are due to **temporal evolution** of the statistical structure of the data.

Regarding the case study considered in this work, the main goal for the temporal analysis of the wind speed data using SODCC is the identification of atypical periods, that can be interesting from climatological point of view. Regarding the spatial evolution, the proposed approach is able to estimate different wind energy production patterns in the area of study. The joint spatio-temporal analysis proposed leads to a robust method for wind resource analysis at different scales (e.g. synoptic), useful to evaluate and compare production in different wind farms in terms of resource evolution in time and space.

2.4. Cluster size probability

One of the first results that can be analyzed from the cluster configurations obtained for each t_I is the cluster size probability $P_{\text{cluster}}(t_j)$. Each $P_{\text{cluster}}(t_j)$ is obtained as following: (i) for all clusters obtained in the independent simulations performed for t_j is calculated the cluster size; (ii) the cluster sizes are assembled in a histogram with bin centers being all possible values; (iii) the histogram is normalized to unitary area, in order to approximate the probability function. Recall that the wind speed data analysis is performed using different independent starting points $t_j \in t_I$, resulting in $|t_I|$ different cluster size probabilities $P_{\text{cluster}}(t_j)$.

The usage of the different $P_{\text{cluster}}(t_j)$ allows both temporal and spatial analysis of the statistics of the data. The temporal evolution can be ascertained by determining differences among the different $P_{\text{cluster}}(t_j)$. Significant differences between $P_{\text{cluster}}(t_j)$ and $P_{\text{cluster}}(t_k)$ are due to significant changes in the statistics of the measured data. Then, they may reveal both change patterns that last long periods of time or isolated t_k where the statistics of the measured data are different compared to the previous and posterior $t_k \in t_I \setminus t_j$. There are multiple options to assess differences between probability distributions, being the KLD one of the most common metrics used in information theory. The KLD will be detailed in the next section.

The different $P_{\text{cluster}}(t_j)$ also gives information about the spatial extent of the data correlations. It is to be expected that multiple cluster sizes will appear for each simulation due to the cluster aggregation strategy of SODCC [21]. The probability of appearance associated to each cluster size will show if the data exhibit either small or large scale spatial extent (e.g. higher probability of large clusters show lead to areas data exhibits correlations at large spatial scales), with no location information.

2.5. KLD : difference between probability distributions

Comparison between different probability distribution can be awkward if there are only little differences between them. The KLD $D_{\text{KL}}(P||Q)$, also called relative entropy, is a metric commonly used in information theory to assess differences between probability distributions [23]. Specifically, it accounts for the information lost if a given probability distribution P is used to approximate a different probability distribution Q. Although often termed as "distance", it is not a literal distance metric as it does not always fulfill the symmetry property, i.e. $D_{\mathrm{KL}}(P||Q) \neq D_{\mathrm{KL}}(Q||P)$. However, this last property has no consequences in the present work, since the objective is to detect differences and not distance between $P_{\mathrm{cluster}}(t_j)$ and $P_{\mathrm{cluster}}(t_k)$, $\{t_j, t_k\} \in t_I$.

The KLD between two cluster size probability distributions $P_{\mathrm{cluster}}(t_j)$ and $P_{\mathrm{cluster}}(t_k)$ is defined as

$$D_{\mathrm{KL}}(t_j, t_k) = \sum_{N_i=N_{\mathrm{min}}}^{N_{\mathrm{max}}} P_{\mathrm{cluster}}(t_k|N_i) \mathrm{log} \frac{P_{\mathrm{cluster}}(t_k|N_i)}{P_{\mathrm{cluster}}(t_j|N_i)} \tag{2}$$

By calculating this metric for all possible combination $\{t_j, t_k\}$, it is possible to quantitatively compare each cluster size probability distribution with the ones obtained from independent simulations with different starting point.

2.6. Node to cluster size probability

In order to further analyze the spatial extent of the data statistics it is necessary to calculate the node to cluster size probability $P_{\mathrm{station}}(t_j, N_i)$, namely the probability of each node to belong to a cluster of size N_i for a given t_j.

The main benefit of $P_{\mathrm{station}}(t_j, N_i)$ is that the representation of each node's probability over a map using color gradient reveals the location of nodes with similar spatial correlation extent. By analyzing different cluster sizes, different statistical behaviours can be identified at different spatial scales. Moreover, the analysis of these results over multiple t_j gives a better insight over over the temporal evolution of the data statistics, as location is also included.

Although this result is very useful for spatial analysis, its main drawback is the large number of possible $P_{\mathrm{station}}(t_j, N_i)$ and that the selection of specific N_i for visualization can be an arduous problem.

3. Experiments and Results

3.1. Dataset

In the experiments carried out, wind speed at 10 m data obtained from the ERA–Interim dataset [19] is considered. The analysis area is bounded by parallels $35°N$ and $44°N$ and by meridians $10°W$ and $6°E$, demarcating the Iberian Peninsula and the Balearic Islands. The grid is set to $0.5° \times 0.5°$, making a total of $N = 627$ nodes in the network, placed in a regular grid as shown in Figure 2. Given this node locations, the distance resolution is approximately $56km$.

Wind speed data for January is considered, in representation for the most windy season in the area of interest [24, 25], from 1979 to 2014, both inclusive. This work only considers *analysis* data with maximum temporal resolution, i.e. 6 hours.

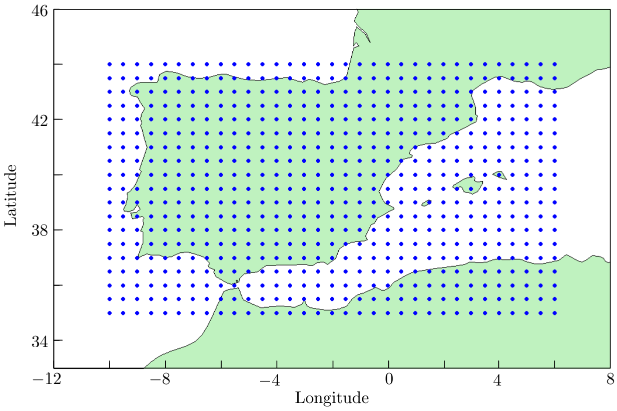


Figure 2: Location of the $N = 627$ reanalysis nodes of the network (blue dots) used in the present work.

3.2. Experiments

In this work, both spatial and temporal correlation of the wind speed data are studied. More precisely, the data of each particular year (between 1979 and 2014) is independently analyzed, giving a total of 36 elements for t_I:

$$t_I \in \{\text{first time instant of } j \mid j = 1979, \ldots, 2014\} \tag{3}$$

For each $t_j \in t_I$, 5000 independent simulations are performed, applying the SODCC algorithm to the wind speed dataset of the $N = 627$ nodes. Such experiments allow the analysis of the clustering distribution probabilities with independence of the initialization of the clustering algorithm. As previously mentioned, the output of each independent SODCC realization is the set of clusters, formed by nearby nodes, where the signal and the noise subspace are separable and it is possible to solve the principal components that explain a minimum of 90% of the variance.

As results, information about the statistical distribution of the cluster sizes is obtained, which is useful to detect major changes in the temporal evolution of the data statistics. The study also focuses on the probability of a given node to belong to a cluster of a given size, which will allow refining the final conclusions regarding both the temporal and spatial changes of the data statistics.

Given the regular grid of the nodes (see Figure 2), it is expected heavy edge effects in the border of the studied area. Hence, the conclusions of this work will be only based on results from the area bounded by parallels $36°N$ and $43°N$ and by meridians $9°W$ and $5°E$.

3.3. Results

3.3.1. Basic cluster statistics

The first results presented in this work are basic statistics over the clusters obtained. On average, there are $265e3$ clusters for each t_j, spanning sizes between $N_{\min} = 3$ and $N_{\max} = 53$, confirming the sufficient statistical representativeness of the results.

The average cluster size, plotted in Figure 3, shows high dispersion for each t_j simulation. Although the average cluster size gives some information about the non-stationarity of the

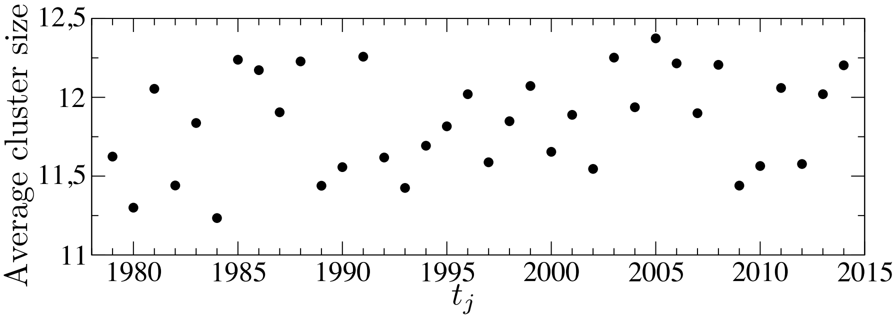


Figure 3: Average cluster size obtained for each $t_j \in t_I$.

wind speed data, it is only a superficial analysis. Previous works shown that the expected cluster size distribution follows a multimodal asymmetric shape [20], where central tendency statistics encapsulate little information.

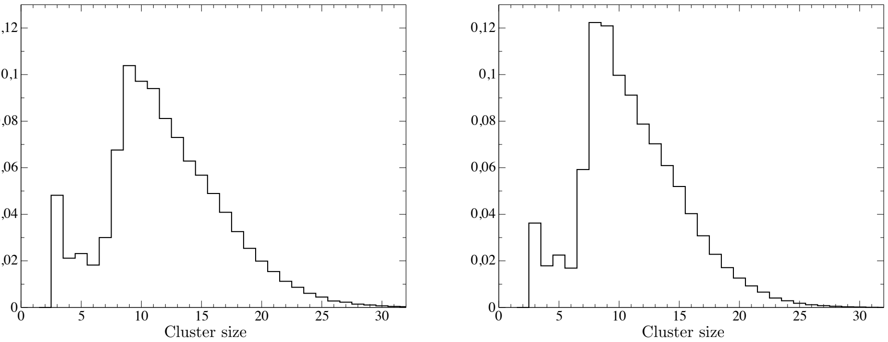
As it can be seen in the following section, the multimodality of the cluster size distribution gives rise to a high variance (second moment statistics) for the average cluster size as soon as the main modes experience even small changes. However, it is possible to identify years with low average cluster size (below 12) as the ones with smaller surface scale in the spatial correlation. It will be shown that the years with lower average cluster size exhibit localized spatial features compared to those with higher average cluster size.

3.3.2. Cluster size probability

In order to assess the temporal evolution of the wind speed data correlations, the cluster size probability distribution for each t_j is considered, namely $P_{\text{cluster}}(t_j)$. Figure 4 shows a sub-set of the obtained distributions. These plots show multi-modal asymmetric distributions where two power-law functions with different origin determine the frequency of each cluster size [20]. These power-law functions emerge because the final cluster configuration of SODCC is based on cluster fusions, a process that follows the *ansatz* of dynamic scaling for aggregation of clusters [26].

The usage of plots like the ones represented in Figure 4 reveals limited information, mainly due to the arduous task to visually compare all of the histograms. Thus, Figure 5 shows all the obtained distributions in a more compact form, using a 3D-bar representation and color gradient for the bar height. The orientation of the figure is chosen to better visualize the most representative differences over the time, such as different slope of the second power-law function.

From Figure 5 it is possible to perceive a quasi-periodic behaviour between the results obtained from different starting points t_j. Figure 6 shows the 3D-bar using a frontal view, where the top of the probability bars representing equal cluster sizes are connected by lines. This figure highlights two key cluster sizes $N_i = 7$ and $N_i = 16$, that show the dichotomy between small size and large size clusters (e.g. if the SODCC clustering based on the measured data leads to many small clusters, there are less large clusters). The quasi-periodic behaviour is better perceived with this representation, particularly for larger cluster sizes. The selection of these key cluster sizes is to aid the visualization, but this



(a) $t_j = 1985$

(b) $T_j = 1986$

Figure 4: Cluster size distribution resulting from the SODCC clustering of the wind speed data from the $N = 627$ nodes for $t_j = 1985$ and $T_j = 1986$.

dual behaviour is apparent for any pair of small and large cluster sizes; see for example the similar trend obtained $N_i \geq 16$.

From the analysis of previous figures, the quasi-periodicity observed reveals the existence of two different patterns in the results, depending on the starting point of the simulation. Then, it is possible to discern between the two cases $t_j \in t_I^*$ and $t_j \in t_I^* \setminus t_I$, being

$$t_I^* \in \{1986, 1988, 1994, 1996, 2001, 2003, 2008, 2010, 2014\} \tag{4}$$

In order to confirm this quasi-periodic behaviour, the differences between the cluster size distributions have been accurately quantified by using the KLD metric. Figure 7 shows $D_{\mathrm{KL}}(t_j, t_k)$ calculated for $t_k \in t_I$ and with a clear separation between the two detected patterns, meaning $t_j \in t_I^*$ and $t_j \in t_I^* \setminus t_I$. This dissociation is useful as $D_{\mathrm{KL}}(t_j \in t_I^*, t_k)$ results in significantly higher values with particular trends that can blur the analysis. This representation accentuates the differences between the distributions with different starting points t_j, and all the cases where $t_j \in t_I^*$ can be easily spotted as their Kullback-Leibler divergence is significantly higher. The exception is $t_j = 2008$ and its particular behaviour will be discussed in Section 4.

3.3.3. Node to cluster size probability

In the previous section it has been shown that there are two different patterns for the cluster size probability functions, due to the different cluster statistics. Following, it is shown how this differences are linked to the spatial and temporal correlations of the wind speed data. For this purpose, the node to cluster size probability $P_{\mathrm{station}}(t_j, N_i)$, has been calculated, namely the probability of a given node to be associated with a cluster of a certain size N_i for the starting point of the simulation t_j.

Figures 8, 9 and 10 represent over a map $P_{\mathrm{station}}(t_j, 7)$, $P_{\mathrm{station}}(t_j, 8)$ and $P_{\mathrm{station}}(t_j, 16)$, respectively. These three cluster sizes are selected for the representation of $P_{\mathrm{station}}(t_j, N_i)$ as they are the ones that illustrate the dichotomy between small size and large size clusters, and so they facilitate a better analysis of the extent of the spatial correlations. In each subfigure,

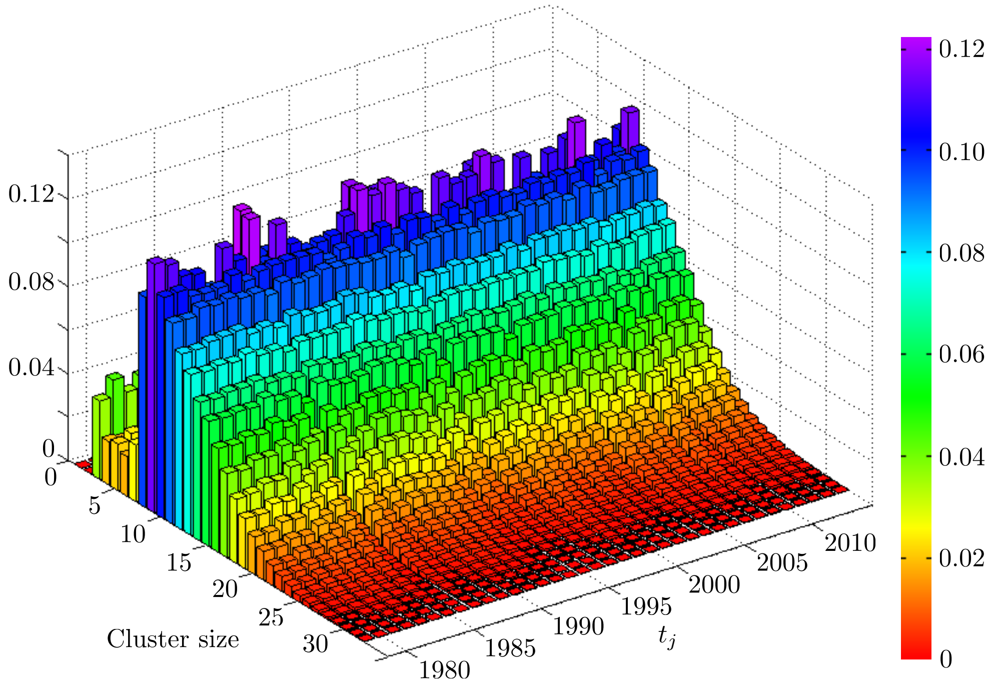


Figure 5: Compact representation of the cluster size distributions resulting from the SODCC clustering of the wind speed data for the $N = 627$ nodes and for all values of $t_j \in t_I$ (see Eq. (3)), represented as a 3D-bar plot. The bar height is indicated with the color gradient.

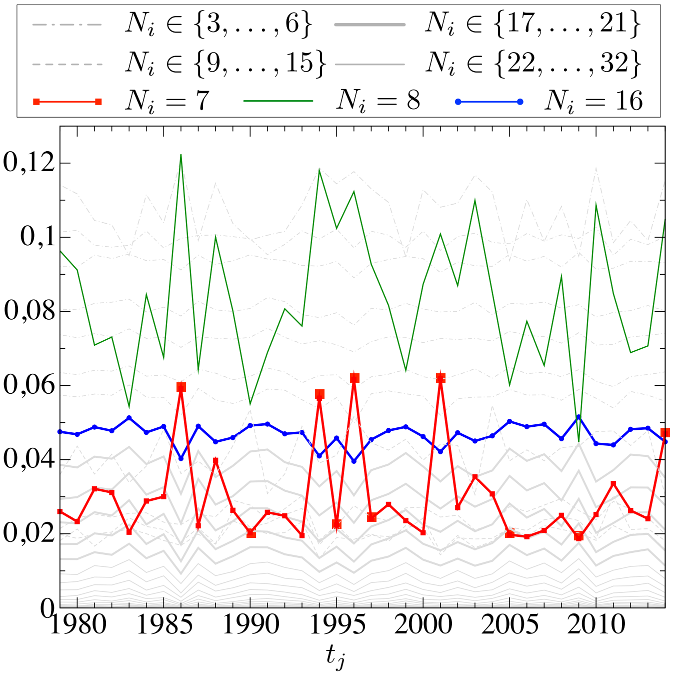
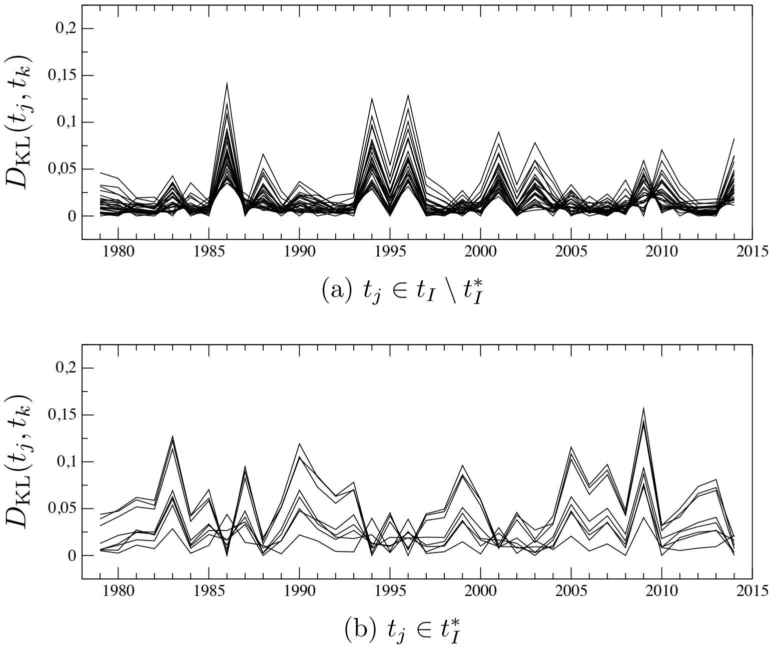


Figure 6: Frontal view of Figure 5. Each line of this figure connects the top of the probability bars for equal cluster sizes. The dichotomy between small and large clusters and their quasi-periodic behaviour is highlighted by the two key cluster sizes $N_i = 7$ and $N_i = 16$.



(a) $t_j \in t_I \setminus t_I^*$

(b) $t_j \in t_I^*$

Figure 7: Kullback-Leibler divergence $D_{\mathrm{KL}}(t_j, t_k)$ for $t_k \in t_I$ and separating between the two patterns detected (a) $t_j \in t_I^* \setminus t_I$ and (b) $t_j \in t_I^*$.

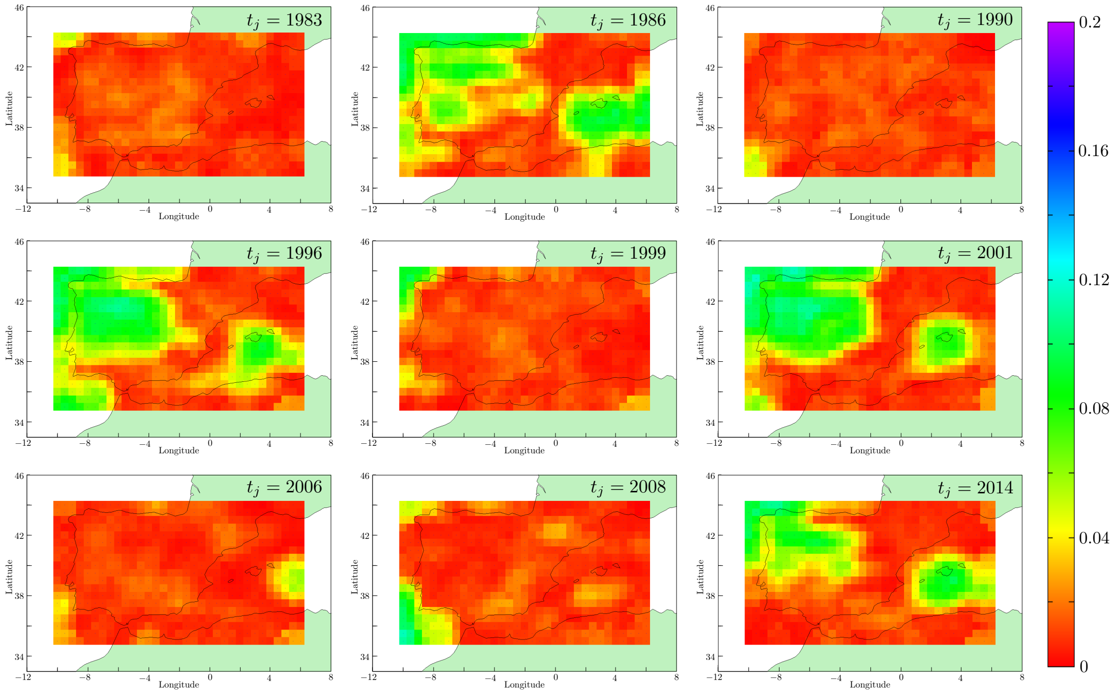


Figure 8: Node to cluster size probability for $N_i = 7$ and multiple values of t_j, i.e. $P_{\mathrm{station}}(t_j, 7)$.

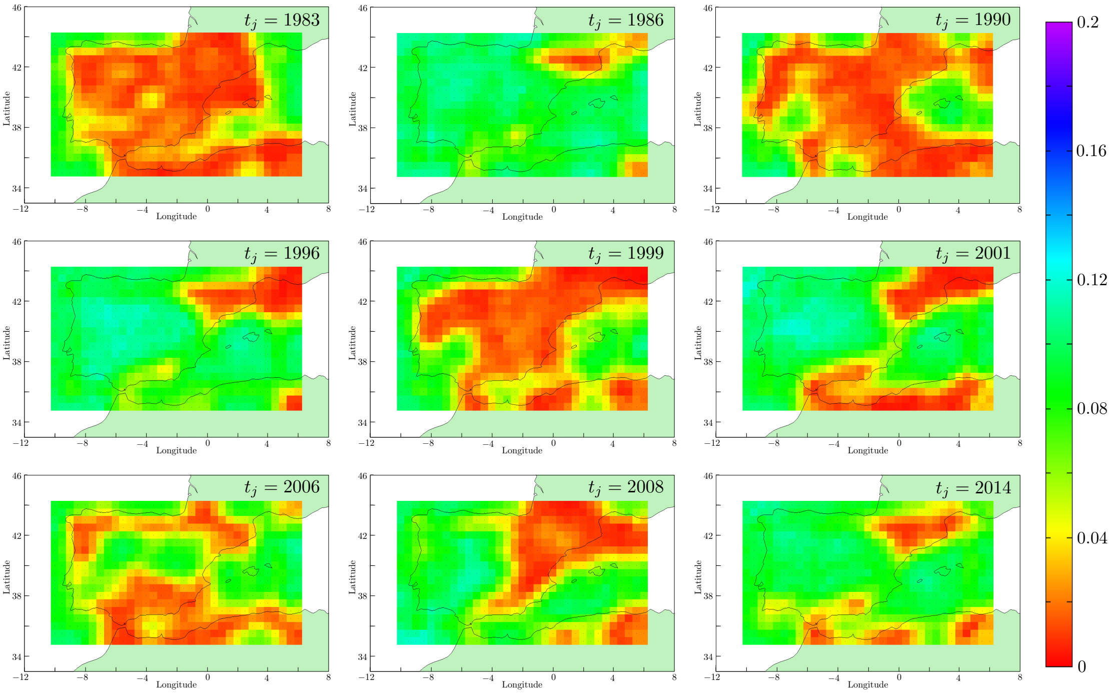


Figure 9: Node to cluster size probability for $N_i = 8$ and multiple values of t_j, i.e. $P_{\text{station}}(t_j, 8)$.

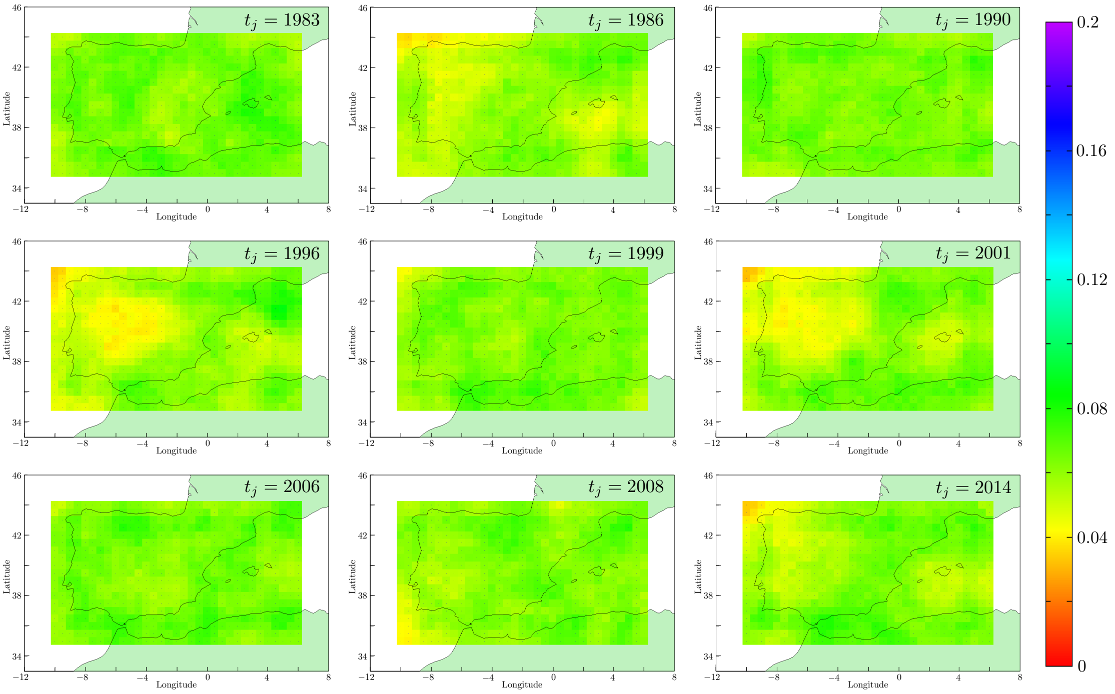


Figure 10: Node to cluster size probability for $N_i = 16$ and multiple values of t_j, i.e. $P_{\text{station}}(t_j, 16)$.

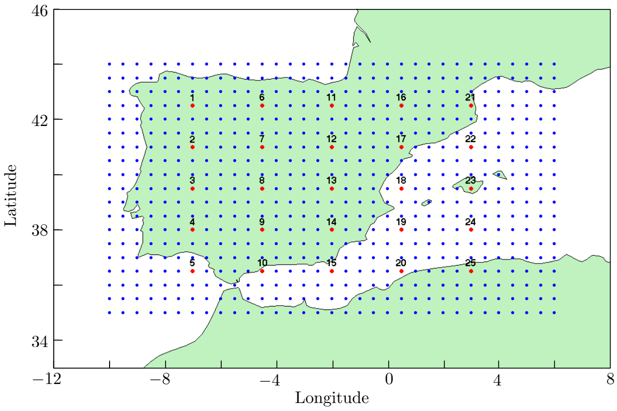


Figure 11: Location and identification of the 25 nodes of the network (red dots) used for the data correlation calculations.

the node location is marked in a map and, for each node, the value of $P_{\text{station}}(t_j, N_i)$ is represented using a color gradient. Note that all figures use the same range for the colorbar, meaning that the same color represent the same probability value.

From these figures it is also straightforward to observe the two patterns identified in our previous results. For example, for $t_j \in t_I^*$ it is possible to clearly identified areas with higher probability than the nodes belong to clusters of sizes $N_i = 7$ and $N_i = 8$. Sharing similar value for $P_{\text{station}}(t_j, N_i)$ indicate that the wind speed data measured by those nodes has similar statistics and, therefore, similar temporal and spatial correlation. Moreover, it is expected that nodes belonging to smaller clusters, their probability to belong to larger clusters is lower, and this behaviour is also confirmed in Figures 8, 9 and 10.

Finally, the most important issue that can be observed in these figures is the following. The clusters obtained for each of the two sets of initialization points ($t_j \in t_I^*$ and $t_j \in t_I \setminus t_I^*$) share much more than similar cluster size probability, obtained without location information. These clusters are formed in the same areas, strengthening the link between the t_j belonging to each of the two sets and confirming the quasi-periodical behaviour already observed in previous results.

3.3.4. Data correlation

In order to validate the previous results and hence the wind speed data analysis method proposed in this work, the correlation between wind speed data measured is now calculated. In this case the 25 sensor nodes plotted in Figure 11 are considered, selected such as they form a regular grid covering all the relevant regions and away from the borders of our analysis area. From this analysis it is expected that nodes located in regions previously identified to share similar statistics will also have higher correlation, compared to nodes located in regions with different node to cluster size probability.

Figure 12 represents the values of the modulus of the correlation between all possible pairs of nodes for multiple values of t_j, using color gradient and considering a threshold of 0.8. Note that the same selection of t_j is considered, as in Figures 8, 9 and 10. From this

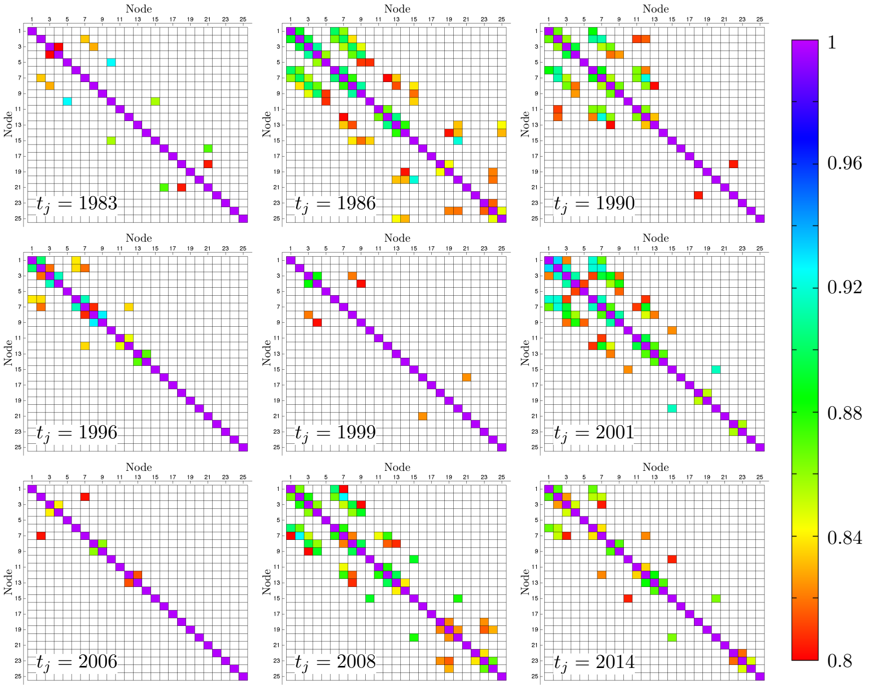


Figure 12: Modulus of the correlation between pairs of nodes represented in Figure 11.

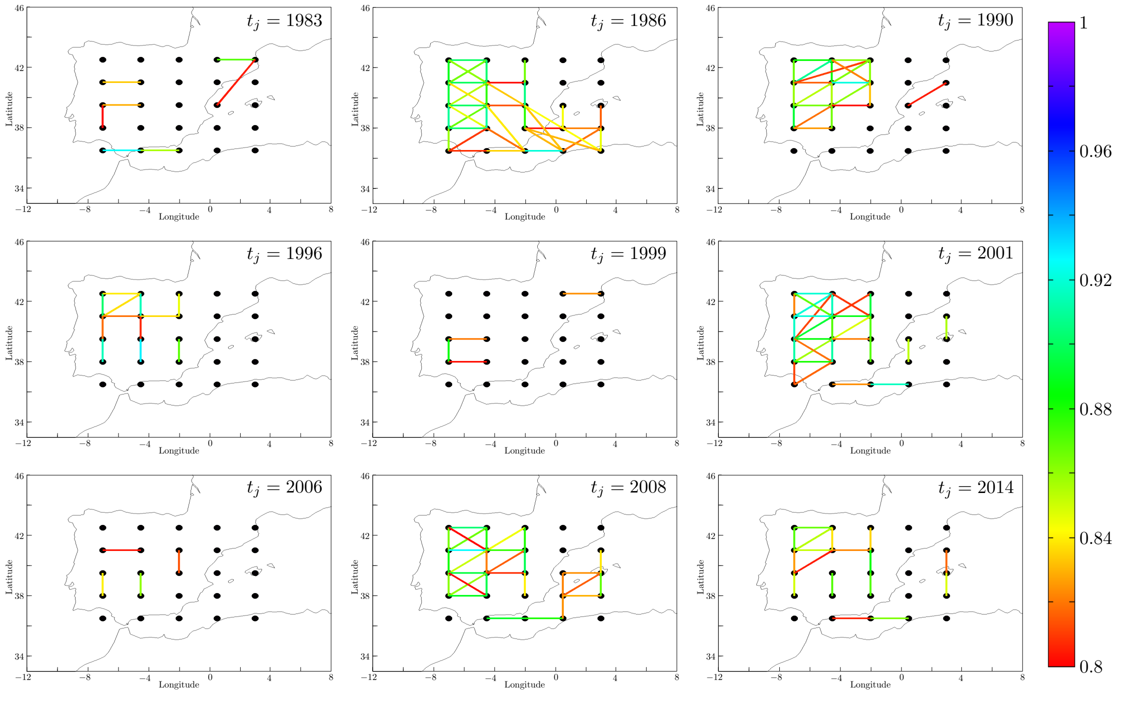


Figure 13: Modulus of the correlation between pairs of nodes represented in Figure 11, shown over a map.

figure the existence two different correlation patterns is confirmed, depending on whether $t_j \in t_I^*$ or $t_j \in t_I \setminus t_I^*$, with significant data correlation between a greater amount of nodes for the case $t_j \in t_I^*$.

Figure 13 shows the modulus of the correlation between the 25 nodes over a map for multiple values of t_j, using color gradient and considering a threshold of 0.8. From this representation it can be observed that the areas that exhibit higher data correlation are similar to the ones that result in similar values of $P_{\text{station}}(t_j, N_i)$ and with higher values of $P_{\text{station}}(t_j, 7)$ and $P_{\text{station}}(t_j, 8)$. This fact confirms that the clusters obtained from SODCC are formed depending on the second order statistics of the measured data and, moreover, clusters with smaller size appear in areas with higher data correlation.

Data analysis based on linear correlations has been used in order to validate the analysis method presented in this work that is based on the SODCC algorithm. The analysis based on linear correlations it is able to show some of the multiple relations present in the data statistics, but has a limited use. On the other hand, the main benefit of using the analysis method presented in this work is that relations between the data can be analyzed at different temporal and spatial scales, for example determined by the value of N_i when the $P_{\text{station}}(t_j, N_i)$.

4. Discussion

This section presents the discussion of the results obtained from the temporal evolution of both the cluster size probabilities $P_{\text{cluster}}(t_j)$ and the node to cluster size probabilities $P_{\text{station}}(t_j, N_i)$.

The main feature of the SODCC algorithm is the search for similarity of second order statistics. Then, the explained variance of the spatial features are accounted for in the peaks of the $P_{\text{cluster}}(t_j)$ distribution. In [20] it was shown that the cluster size distribution can written as:

$$P_{\mathsf{cluster}}(t_j; \tau) = \kappa \times \tau^{-w} \left(g(N_i, \hat{d}_{\mathsf{min}}) + g(N_i, \hat{d}_{\mathsf{max}}) \right) \tag{5}$$

where w is the decay exponent due to the power-law behaviour of the aggregation process, and

$$g(x, x_0) = \theta(x - x_0)(x - x_0)^{-\tau} \tag{6}$$

$\theta(x_0)$ is the Heaviside step-function and

$$\kappa = \frac{\tau - 1}{(N - \hat{d}_{\mathsf{min}})^{1-\tau} + (N - \hat{d}_{\mathsf{max}})^{1-\tau} - 2} \tag{7}$$

is a distribution normalization constant. The $\hat{d}_{\mathsf{min}}$ and $\hat{d}_{\mathsf{max}}$ refer, respectively, to the smallest and largest number of principal components (or signal eigenvalues) found in the set of achieved clusters. Thus, a two-peaked distribution with a sharp exponential decay after each of the peaks as the cluster size increases can be expected. What Eq. (5) intuitively shows is the identification of both small and large patches of nearby spatial locations (respectively small and large clusters) that are tightly related (in terms of variance of wind speed). Those are the two peaks of the distribution (see Figure 4). After those peaks (in-between and after the largest one), spatial areas that occur with the same exponentially decreasing probability. Thus, the peaks of the cluster size distribution identify the smallest and largest areas where the explained variance can be ascertained in terms of a fewer number of eigenvalues than evaluated sites.

Due to the fact that the exponential decay of the $P_{\text{cluster}}(t_j)$ after the largest peak has a characteristic constant τ (that does not depend on the characteristics of the data but on the dynamics of the self-organization of clusters), any variation on the position of the largest peak will be reflected in noticeable variations in the exponential decay. As it can be seen in the 3D stacked representation of $P_{\text{cluster}}(t_j)$ in Figure 5, an oscillating pattern on the side of the "mountain range" can be identified. The pattern is made more apparent in a 2D projection of iso-levels of $P_{\text{cluster}}(t_j)$ (iso-cluster size levels) in time (see Figure 6). A surge in the probability of occurrence for clusters of size $N_i = 8$ for a given year, is followed by a sharp decay of occurrence in the largest clusters and vice versa. The pattern is shown to evolve smoothly with time. As the $P_{\text{cluster}}(t_j)$ for different t_j have been calculated independently, the smooth evolution of $P_{\text{cluster}}(t_j)$ in time must be due to the wind speed data in the different years. The results for the evolution in time for clusters of size $N_i = 8$ show peaks in years 1986, 1988, 1994, 1996, 2001, 2003, 2008, 2010 and 2014, which are the elements of the set T_I^*.

In order to quantify the differences in $P_{\text{cluster}}(t_j)$ for different t_j, the KLD (Eq. (2)) is used and the results are presented in Figure 7. The two families of patterns identified are the following: Figure 7(a) represents the KLD distance from $P_{\text{cluster}}(t_j)$ in years with low probability of forming clusters of size 7 and 8; In Figure 7(b) the KLD uses the "peaked" years mentioned above (t_I^*). It can be clearly seen that two distinct patterns emerge, thus supporting the hypothesis that two spatio-temporal behaviors of spatial similarity in wind speed can be identified in the Iberian Peninsula through time. As the spatial node to cluster size probability is plotted for a subset of both normal and anomalous years (t_I^*) for cluster sizes $N_i = 7$ and 8 (see Figures 8 and 9) an emergence of local patterns in the NW of the Iberian Peninsula and the Balearic Islands is apparent (green patches) for peaked years (1986, 1996, 2001 and 2014). These spatial patterns show evidence of a tight correlation in the wind speed (i.e. explained variance of the wind speed). Figure 10 shows the "complementary" figure to the previous one, i.e., where small clusters ($N_i = 7$ and 8) are bound to be appear, large clusters ($N_i = 16$) do not appear and vice versa.

However, there is an apparent exception to the behaviour described above, which is year $t_j = 2008$. By the apparent behaviour at the cluster size at $N_i = 8$ of such year (Figure 9), it shares the same characteristics as years 1986, 1996, 2001 and 2014. These similarities are also observed in the traditional correlation analysis shown in Figures 12 and 13. But at the shorter scale of $N_i = 7$ (Figure 8), the differences between 2008 and the aforementioned set are more than apparent. This difference would indicate that atmospheric dynamics with a shorter time (and length) scale are taking place for all the years included the set of the anomalous years t_I^*, except for $t_j = 2008$ when these dynamics take place at an intermediate level. This would indicate that 2008 is a transitional year between anomalous and normal. This fact is also captured by the lower amplitude of the KLD statistic of year 2008 (Figure

7).

The power of the present method is apparent for the results that can be obtained at limited computational complexity. Its computational complexity is dependent on the average cluster sizes that emerge and not on the number of total sites evaluated [20]:

$$\langle N_i \rangle \sim \frac{2}{\tau - 1} + \frac{2N^{2-\tau}}{2-\tau} \tag{8}$$

where the notation $\langle \cdot \rangle$ indicates the expectation operation. The value of $\langle N_i \rangle$ in the present work is 13 thus, the computational complexity of the method is proportional to 13^2.

As a comparison a similar spatio-temporal analysis is performed, e.g. a cross-correlation analysis with a subset of sites (25 out of 627) in the Iberian Peninsula (shown in Figure 11). The results of the modulus of the cross-correlation matrix for a minimum threshold of 0.8 for the same set of normal and anomalous years are presented in Figure 12. There it is possible to see that only a few of the node pairs posses a cross-correlation value above the threshold. Moreover, Figure 13 shows the projection of a colored line connecting those nodes (with the color representing the cross correlation value) over their corresponding positions in space, a similar pattern as the one shown in Figures 8 and 9, but with much less resolution and a computational complexity of 25^2.

The relevance of the presented result to the wind power generation community are patent. Wind speed is generally regarded and modelled as a Weibull distribution [27, 28]:

$$f(v; A, k) = \frac{k}{A}\left(\frac{v}{A}\right)^{k-1} e^{-\left(\frac{v}{A}\right)^k} \tag{9}$$

where A is the scale parameter of the distribution and k is its shape parameter.

The present method focuses on the clustering of sites with a explained variance of minimal complexity, thus it tends to group into the same cluster sites with similar second order statistics of wind speed. The n-th order statistic of a Weibull distributed variable can be calculated as:

$$\langle v^n \rangle = \int_0^\infty v^n f(v; A, k) dv = A^n \Gamma\left(\frac{n}{k} + 1\right) \tag{10}$$

The average power generated by a wind turbine can be calculated as

$$\langle P \rangle = \frac{1}{2}\rho \int_0^\infty v^3 f(v; A, k) dv = \frac{1}{2}\rho A^3 \Gamma\left(\frac{3}{k} + 1\right) \tag{11}$$

It is possible to calculate the ratio of the third-to-second order statistic of the wind speed as

$$\frac{\langle v^3 \rangle}{\langle v^2 \rangle} = A \frac{\Gamma\left(\frac{3}{k} + 1\right)}{\Gamma\left(\frac{2}{k} + 1\right)} \tag{12}$$

For typical values of shape parameter k of the distribution of wind speed, it is possible to show that the previous ratio is very slowly varying with k, such that

$$1 > \frac{1}{3A}\frac{\langle v^3 \rangle}{\langle v^2 \rangle} > \frac{\sqrt{\pi}}{4}, \text{ for } 1 < k < 2 \tag{13}$$

Thus, for the regions of interest, it is possible to rewrite Eq. (11) as

$$\langle P\rangle \approx \frac{1}{2}\rho k A^3 \Gamma\left(\frac{2}{k}+1\right) = \frac{1}{2}\rho k A \left\langle v^2\right\rangle \tag{14}$$

Therefore, this work shows that, when the SODCC clusters regions of similar variance of wind speed, it approximately clusters regions of similar power output from wind turbines also. This makes the present method of particular interest for wind energy applications and forecasting at very limited computational cost and increased resolution.

5. Conclusions

This paper presents a spatio-temporal analysis method of the wind resource, focused in the Iberian Peninsula. This work also presents a detailed description of the proposed method, that is based on the SODCC algorithm which clusters a network in terms of the measured data statistics. Extensive experiments have been also performed, by applying the SODCC algorithm to reanalysis data in the area of interest for the period 1979 - 2014. Based on the obtained results, two different spatio-temporal patterns of the data statistics depending on the initialization year have been identified. A reasonable hypothesis is that these two patterns are forcefully related to the main driver of wind speed in the Iberian Peninsula in the winter. This is quantified by the difference between atmospheric pressures in Lisbon and Rejkiavik, namely, the North Atlantic Oscillation (NAO). In future works, the relation between the KLD in the Iberian Peninsula and the NAO sequences will be studied.

The relation that has been established in the previous section between the average power of the wind turbine and the variance of the wind speed allows further insight into differences in energy production in shorter time and length scales. As smaller cluster sizes imply faster atmospheric dynamics, it is possible to deduce that small clusters exhibit higher variance in wind speed. This, in turn, would suggest that in zones that belong to small, persistent in time, clusters the power generated by wind turbines should be higher. At the present resolution of 0.5° it is not practical to identify such zones for wind farm positioning as each point is separated more than 50 km. However, in future works this methodology will be used to identify such zones. As it stands, the proposed method has a great potential for the spatio-temporal analysis of offshore wind farms, in which the spatial scales are much more relevant than in the onshore case.

Acknowledgements

ERA-Interim data provided courtesy of the ECMWF.

This work has been partially supported by *Comunidad de Madrid*, under projects number S2013/ICE-2933 and S2013/MAE-2835, and by project TIN2014-54583-C2-2-R of the *Spanish Ministerial Commission of Science and Technology (MICYT).*

Mihaela I. Chidean is supported by the FPU Research Grant AP2012-2981 from the Spanish Ministry of Education, Culture and Sports.

References

[1] GWEC, Global Wind Report 2015, `http://www.gwec.net/publications/global-wind-report-2/global-wind-report-2015-annual-market-update/`, online.

[2] B. D. Cross, K. E. Kohfeld, J. Bailey, A. B. Cooper, The impacts of wind speed trends and 30-year variability in relation to hydroelectric reservoir inflows on wind power in the pacific northwest, PloS one 10 (8) (2015) 1–22.

[3] D. Getman, A. Lopez, T. Mai, M. Dyson, Methodology for clustering high-resolution spatiotemporal solar resource data, National Renewable Energy Laboratory, Technical Report NREL/TP-6A20-63148.

[4] A. Ganske, B. Tinz, G. Rosenhagen, H. Heinrich, Interannual and multidecadal changes of wind speed and directions over the north sea from climate model results, Meteorologische Zeitschrift 25 (4) (2016) 463–478.

[5] D. Jiang, D. Zhuang, Y. Huang, J. Wang, J. Fu, Evaluating the spatio-temporal variation of China's offshore wind resources based on remotely sensed wind field data, Renewable and Sustainable Energy Reviews 24 (2013) 142–148.

[6] L. Yu, S. Zhong, X. Bian, W. E. Heilman, Temporal and spatial variability of wind resources in the United States as derived from the Climate Forecast System Reanalysis, Journal of Climate 28 (3) (2015) 1166–1183.

[7] M. Z. Ibrahim, Y. K. Hwang, M. Ismail, A. Albani, Spatial Analysis of Wind Potential for Malaysia, International Journal Of Renewable Energy Research 5 (1) (2015) 201–209.

[8] F. Santos-Alamillos, N. Thomaidis, S. Quesada-Ruiz, J. Ruiz-Arias, D. Pozo-Vázquez, Do current wind farms in spain take maximum advantage of spatiotemporal balancing of the wind resource?, Renewable Energy 96 (2016) 574–582.

[9] R. Arjmand, M. Rahimiyan, Impact of spatio-temporal correlation of wind production on clearing outcomes of a competitive pool market, Renewable Energy 86 (2016) 216–227.

[10] A. Troccoli, K. Muller, P. Coppin, R. Davy, C. Russell, A. L. Hirsch, Long-term wind speed trends over Australia, Journal of Climate 25 (1) (2012) 170–183.

[11] A. Tascikaraoglu, B. M. Sanandaji, K. Poolla, P. Varaiya, Exploiting sparsity of interconnections in spatio-temporal wind speed forecasting using wavelet transform, Applied Energy 165 (2016) 735–747.

[12] V. M. Gomez-Muñoz, M. Porta-Gándara, Local wind patterns for modeling renewable energy systems by means of cluster analysis techniques, Renewable Energy 25 (2) (2002) 171–182.

[13] A. Kusiak, W. Li, Short-term prediction of wind power with a clustering approach, Renewable Energy 35 (10) (2010) 2362–2369.

[14] A. Di Piazza, M. C. Di Piazza, A. Ragusa, G. Vitale, Environmental data processing by clustering methods for energy forecast and planning, Renewable energy 36 (3) (2011) 1063–1074.

[15] G. Grigoras, F. Scarlatache, An assessment of the renewable energy potential using a clustering based data mining method. case study in romania, Energy 81 (2015) 416–429.

[16] D. Liu, J. Wang, H. Wang, Short-term wind speed forecasting based on spectral clustering and optimised echo state networks, Renewable Energy 78 (2015) 599–608.

[17] E. T. Al-Shammari, S. Shamshirband, D. Petković, E. Zalnezhad, L. Yee, R. S. Taher, Ž. Ćojbašić, Comparative study of clustering methods for wake effect analysis in wind farm, Energy 95 (2016) 573–579.

[18] L. Dong, L. Wang, S. F. Khahro, S. Gao, X. Liao, Wind power day-ahead prediction with cluster analysis of nwp, Renewable and Sustainable Energy Reviews 60 (2016) 1206–1212.

[19] D. P. Dee, S. M. Uppala, et al., The ERA–Interim reanalysis: Configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society 137 (656) (2011) 553–597.

[20] M. I. Chidean, E. Morgado, E. del Arco, J. Ramiro-Bargueño, A. J. Caamaño, Scalable Data-Coupled Clustering for Large Scale WSN, IEEE T Wireless Commun 15 (2015) 4681–4694.

[21] M. I. Chidean, J. Muñoz-Bulnes, J. Ramiro-Bargueño, A. J. Caamaño, S. Salcedo-Sanz, Spatio-temporal trend analysis of air temperature in europe and western asia using data-coupled clustering, Global and Planetary Change 129 (0) (2015) 45 – 55.

[22] G. Xu, T. Kailath, Fast subspace decomposition, IEEE Trans. on Signal Processing 42 (3) (1994) 539–551.

[23] T. M. Cover, J. A. Thomas, Elements of information theory, John Wiley & Sons, 2012.

[24] K. M. Zishka, P. J. Smith, The climatology of cyclones and anticyclones over North America and surrounding ocean environs for January and July, 1950-77, Monthly Weather Review.

[25] M. L. Branick, A climatology of significant winter-type weather events in the contiguous United States, 1982-94, Weather and Forecasting.

[26] T. Vicsek, F. Family, Dynamic Scaling for Aggregation of Clusters, Physical Review Letters 52 (19) (1984) 1669–1672.

[27] B. Saavedra-Moreno, S. Salcedo-Sanz, C. Casanova-Mateo, J. Portilla-Figueras, L. Prieto, Heuristic correction of wind speed mesoscale models simulations for wind farms prospecting and micrositing, Journal of Wind Engineering and Industrial Aerodynamics 130 (2014) 1–15.

[28] H. Goh, S. Lee, Q. Chua, K. Goh, K. Teo, Wind energy assessment considering wind speed correlation in malaysia, Renewable and Sustainable Energy Reviews 54 (2016) 1389–1400.