Electrical Power Fluctuations in a Network of DC/AC inverters in a Large PV Plant: relationship between correlation, distance and time scale

O. Perpiñán[a,b,1,*], J. Marcos[c], E. Lorenzo[a]

[a] *Instituto de Energía Solar, Campus Sur, Carretera de Valencia km. 7 28031 Madrid, Spain.*
[b] *Electrical Engineering Department, EUITI-UPM, Ronda de Valencia 3, 28012 Madrid, Spain.*
[c] *Dpto. Ingeniería Eléctrica y Electrónica, Universidad Pública de Navarra, Edificio Los Pinos, Campus Arrosadia, 31006 Pamplona, Spain.*

Abstract

This paper analyzes the correlation between the fluctuations of the electrical power generated by the ensemble of 70 DC/AC inverters from a 45.6 MW PV plant. The use of real electrical power time series from a large collection of photovoltaic inverters of a same plant is an important contribution in the context of models built upon simplified assumptions to overcome the absence of such data.

This data set is divided into three different fluctuation categories with a clustering procedure which performs correctly with the clearness index and the wavelet variances. Afterwards, the time dependent correlation between the electrical power time series of the inverters is estimated with the wavelet transform. The wavelet correlation depends on the distance between the inverters, the wavelet time scales and the daily fluctuation level. Correlation values for time scales below one minute are low without dependence on the daily fluctuation level. For time scales above 20 minutes, positive high correlation values are obtained, and the decay rate with the distance depends on the daily fluctuation level. At intermediate time scales the correlation depends strongly on the daily fluctuation level.

The proposed methods have been implemented using free software. Source code is available as supplementary material.

Keywords: irradiance and electrical power fluctuation, wavelet analysis, wavelet variance, wavelet cross-correlation.

Nomenclature

a, b, c Coefficients of the exponential decay model.

$\widetilde{\mathcal{D}}_j$ jth level detail of an MRA with a MODWT.

$\gamma_{\tau,XY}(\lambda_j)$ Wavelet cross-covariance of two stochastic processes X_t and Y_t for scale λ_j and lag τ.

$\gamma_{XY}(\lambda_j)$ Wavelet covariance with lag zero.

[*] Corresponding author
Email address: `oscar.perpinan@upm.es` (O. Perpiñán)
[1] ISES member

GCR Ground cover ratio

GRR Ground requirement ratio

λ_j Scale j of a wavelet transform.

MODWT Maximum Overlap Discrete Wavelet Transform

MRA Multiresolution analysis

$\nu^2_{X,j}$ Wavelet variance of the scale λ_j.

$\rho_{\tau,XY}(\lambda_j)$ Wavelet cross-correlation of two stochastic processes X_t and Y_t for scale λ_j and lag τ.

$\rho_{XY}(\lambda_j)$ Wavelet correlation with lag zero.

SDF Spectral density function.

σ^2_X Variance of the time series X_t.

$\widetilde{\mathcal{S}}_{J_0}$ J_0th level smooth of an MRA with a MODWT.

$\widetilde{\mathcal{V}}$ $N \times N$ real-valued MODWT scaling matrix.

$\widetilde{\mathbf{W}}$ N dimensional vector of MODWT coefficients.

$\widetilde{\mathcal{W}}$ $N \times N$ real-valued MODWT wavelet matrix.

WT Wavelet transform

$\mathbf{X}$ N dimensional vector containing a real-valued time series.

X_t Real-valued time series.

1. Introduction

Short-term fluctuations in the power generated by large PV plants due to changes in cloud cover can negatively affect utility grid stability and reliability. This fact, together with the high levels of penetration achieved by PV power generation sector over the last few years, has alerted grid operators in some countries, promoting research initiatives to study these fluctuations. The main efforts are trying to quantify the smoothing effect in fluctuations registered not only in one PV plant, but also in an ensemble of geographical dispersed large PV plants and to understand the temporal and spatial relationship between fluctuations.

A bibliographic review reveals the growing concern about this problem. Otani et al. [1] showed that for distances between the locations greater than 5 km, observed daily irradiances fluctuations with 1 min resolution are essentially uncorrelated. Later on, the same authors proposed a method to estimate the largest power fluctuation during a month as the product of the standard deviation fluctuation by a so called “largest fluctuation coefficient” [2]. Wiemken et al. [3] worked with one year of 1 min data from 100 PV sites (totalling 243 kWp) spread over Germany. They observed that, at that scale, power fluctuations of the normalised ensemble power are reduced to 10%. Hoff et al. [4] perform a mathematical analysis which quantifies the variability reduction in power fluctuation from a fleet of PV systems, ranging from individual systems to a set of distributed systems. A relationship between the variance of the fluctuations of a single PV plant and an ensemble is suggested. Subsequently, they introduce a novel approach to estimate the maximum short-term output variability from an arbitrary fleet of PV systems [5], proposing the necessity of real power data to test and validate the models.

Some contributions to the field by the authors have brought an empirical expression obtained via real measurements to compare the fluctuation attenuation because of both the size and the number of PV plants grouped [6, 7]. Observed short-term fluctuations are essentially uncorrelated for distances between PV plants over 6 km. However, the authors remark the need to examine the smoothing effect below that distance.

On the other hand, the wavelet transform (WT) is used in [8] to show that the irradiance and power time series are nonstationary processes whose behaviour resembles that of a long memory process. The combination of a wavelet variance analysis with the long memory spectral exponent shows again that a PV plant behaves as a low-pass filter.

This paper contributes to previous findings with the analysis of experimental data from a large PV plant. The data comprises almost 2 years of irradiance and wind speed time series from a meteorological station, and electrical power time series from a set of 70 inverters. Distances between inverters range from 220 meters to 2.8 kilometers. We analyze the correlation between the fluctuations of the electrical power from each inverter at different time scales and distances, and the connection between the daily level of global irradiance fluctuations and the correlation between the electrical power from the inverters across different time scales and distances.

It must be highlighted the importance of the use of real electrical power time series from a large collection of photovoltaic inverters of a same plant. To overcome the absence of such data, previous research have proposed simplified models with several assumptions:

- A PV plant is modelled as a virtual network composed of identical one-dimensional PV installations [4]. Therefore, electrical mismatch and shadow effects in the PV generator, and performance differences between inverters are not considered, although these effects alter the correlation between power fluctuation time series.

- The electrical power from virtual networks is approximated as the product of the plane-of-array irradiance at one of the systems of the network and a constant factor. The model is further simplified with the use of global horizontal irradiance instead of the irradiance incident on the plane of the system [4, 9]. It must be noted that, at least on a daily basis, the variability of the effective irradiation incident on tracking planes has been reported to be higher than the variability of irradiation on the horizontal plane [10, 11].

- Clouds do not change but travel at constant one-dimensional speed. Cloud speeds are estimated from satellite image analysis with an operational frequency of one per hour. High frequency estimations are obtained via linear interpolation [5, 12].

The use of real electrical time series circumvents the need for these simplifications that generate uncertainty and imposes a limit on the confidence in the output of these models.

Other important contributions of this paper are the examination of the relation between the daily irradiance fluctuation level and several meteorological features, and the unsupervised classification (or clustering) of a collection of days in three different categories according to the fluctuation level. Instead of using a parametrical approach, the features structure is not specified a priori but is instead mostly determined from data. This paper proposes a collection of features to describe the data relying on very few assumptions. With the results of the clustering, the correlation at different time scales and distances is examined in relation with the clustering results.

Finally, it must be noted that the methods have been implemented using free software. Source code is available as supplementary material (section 3.1).

This paper is organized as follows. In section 2 the PV plant and the data acquisition system is detailed. The mathematical formulation and discussion of the proposed methods is provided in the section 3: the wavelet transform is summarized in the section 3.2, the feature analysis and the clustering procedure are described in the section 3.4, and the wavelet correlation is the subject of the section 3.3. The results of these methods are analyzed in section 4 and main conclusions of this research are provided in section 5.

2. PV plant

The experimental data analysed in this paper belongs to a large PV plant situated at Amareleja (Portugal) and owned by Acciona Energía. The plant is spread over 250 Ha and incorporates 2,520 solar trackers, ranging from 17.7 to 18.2 kWp, summing up a total generator power of 45.6 MWp (38.5 MW MV/HV transformer power). The corresponding ground cover ratio, GCR[2], of the plant is 0.162 (its inverse, the ground requirement ratio, GRR, is 6.17). Each tracker is tilted 45° from the floor and is mounted on a vertical-axis tracker (azimuth) paralleling the sun east–west motion. The plant is divided in 70 generators. Each generator consists of 36 trackers feeding a 550 kW DC/AC inverter. The distances between them range from 220 meters to 2.8 kilometers. Each inverter has its own 20 kV transformer and the whole PV plant feed power to the 66 kV grid by a 44 MVA transformer. This PV system was started up in December 2008 and the annual estimated production is over 93 GWh.

The PV plant has been equipped with an intensively monitoring system, which has provided us the experimental data base. Every 5 seconds the power generated by each inverter is synchronously recorded. Simultaneously, meteorological data is also recorded by the PV plant weather station, which provides a measurement of global irradiance and wind speed among others. Data recording started in May 2010 and is still undergoing.

3. Methods

The time-dependent correlation between the electrical power time series of the inverters in different wavelet scales is calculated for each day of the data set (figure 1):

1. The wavelet coefficients of each of the 70 time series for the scales 1 to 9 are calculated (section 3.2.1).
2. The scale-dependent correlations between the power time series of each of the possible combinations between the 70 inverters of the plant is estimated with the wavelet correlation (section 3.3). This results in a set of nine wavelet correlation matrices with 70 rows and columns for each day of the data set.

In order to relate the wavelet correlation with the fluctuation level, the data set is partitioned in three different groups according to the daily fluctuation behaviour of the global irradiance with the meteorological measurements registered by the PV plant weather station (section 3.4, figure 2). Each day is characterized with a set of features:

1. The global irradiation on the horizontal plane and the clearness index.
2. The wavelet variance for different scales (section 3.2.2).
3. The mean, minimum and maximum wind speed.

After a feature selection procedure (section 3.4.2), and with a matrix of transformed values (section 3.4.3), the PAM algorithm (section 3.4.1) divides the set of daily values in three different clusters of low, medium and high fluctuation levels (section 3.4.4).

[2] GCR is defined as the ratio between PV array area to total ground area.

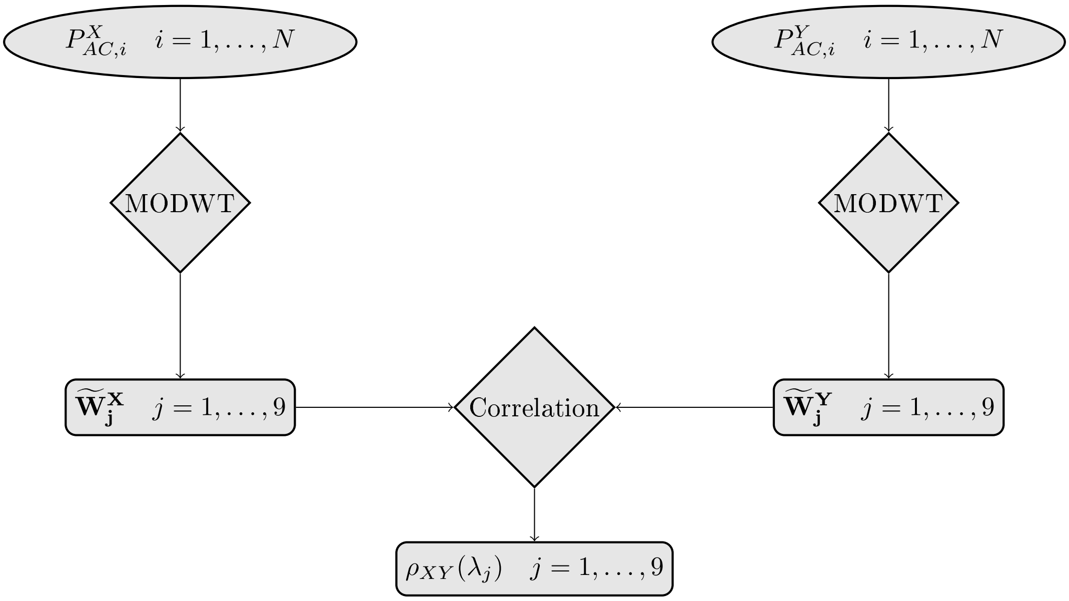

$P^X_{AC,i} \quad i = 1, \ldots, N$
$P^Y_{AC,i} \quad i = 1, \ldots, N$
MODWT
MODWT
$\widetilde{\mathbf{W}}^{\mathbf{X}}_{\mathbf{j}} \quad j = 1, \ldots, 9$
Correlation
$\widetilde{\mathbf{W}}^{\mathbf{Y}}_{\mathbf{j}} \quad j = 1, \ldots, 9$
$\rho_{XY}(\lambda_j) \quad j = 1, \ldots, 9$

Figure 1: Wavelet correlation algorithm. $P^X_{AC,i}$ and $P^Y_{AC,i}$ are the *i*-th elements of the electrical power time series from inverters X and Y, respectively. $\widetilde{\mathbf{W}}^{\mathbf{X}}_{\mathbf{j}}$ and $\widetilde{\mathbf{W}}^{\mathbf{Y}}_{\mathbf{j}}$ are the correspondent wavelets coefficients of scale λ_j, and $\rho_{XY}(\lambda_j)$ is the wavelet correlation of the inverters X and Y at wavelet scale λ_j.

3.1. *Software*

The methods described below have been implemented using the free software environment R [27] and several contributed packages, namely: `zoo` [28] for the time series, `wmtsa` [29] for the wavelet analysis, `solaR` [30] for the clearness index, `sp` [31], `lattice` and `latticeExtra` [32, 33] for displaying the results, `cluster` [34] for the clustering analysis and `car` [35] for the Box-Cox functions. Throughout this section several footnotes are included to provide more information about the connection between methods and code. The source code is available at `https://github.com/oscarperpinan/wavCorPV`.

3.2. *Wavelet analysis*

The analysis of solar irradiance and electrical power time series with the wavelet transform can be found in a variety of research papers [8, 9, 13]. The wavelet transform is a filtering procedure which carries out a multiresolution analysis (MRA) of a time series, where each of the decompositions of the analysis is a representation of the original signal with a different temporal scale. This non-parametric analysis overcomes the restrictions imposed when using classical time series analysis: for example, AR, MA, ARMA models and spectral analysis, are based on the premises of stationarity and short memory process, conditions which are not fulfilled by solar irradiance time series [8]. Besides, this tool improves the examination of fluctuations in a time series: for example, it is common to find reports with fluctuations calculated as a simple substraction between two samples, with time intervals chosen arbitrarily, without relation with data.

Here we use the wavelet variance as a measure of fluctuation of solar irradiance in different scales (section 3.4), and the MODWT of the power time series of each inverter to estimate

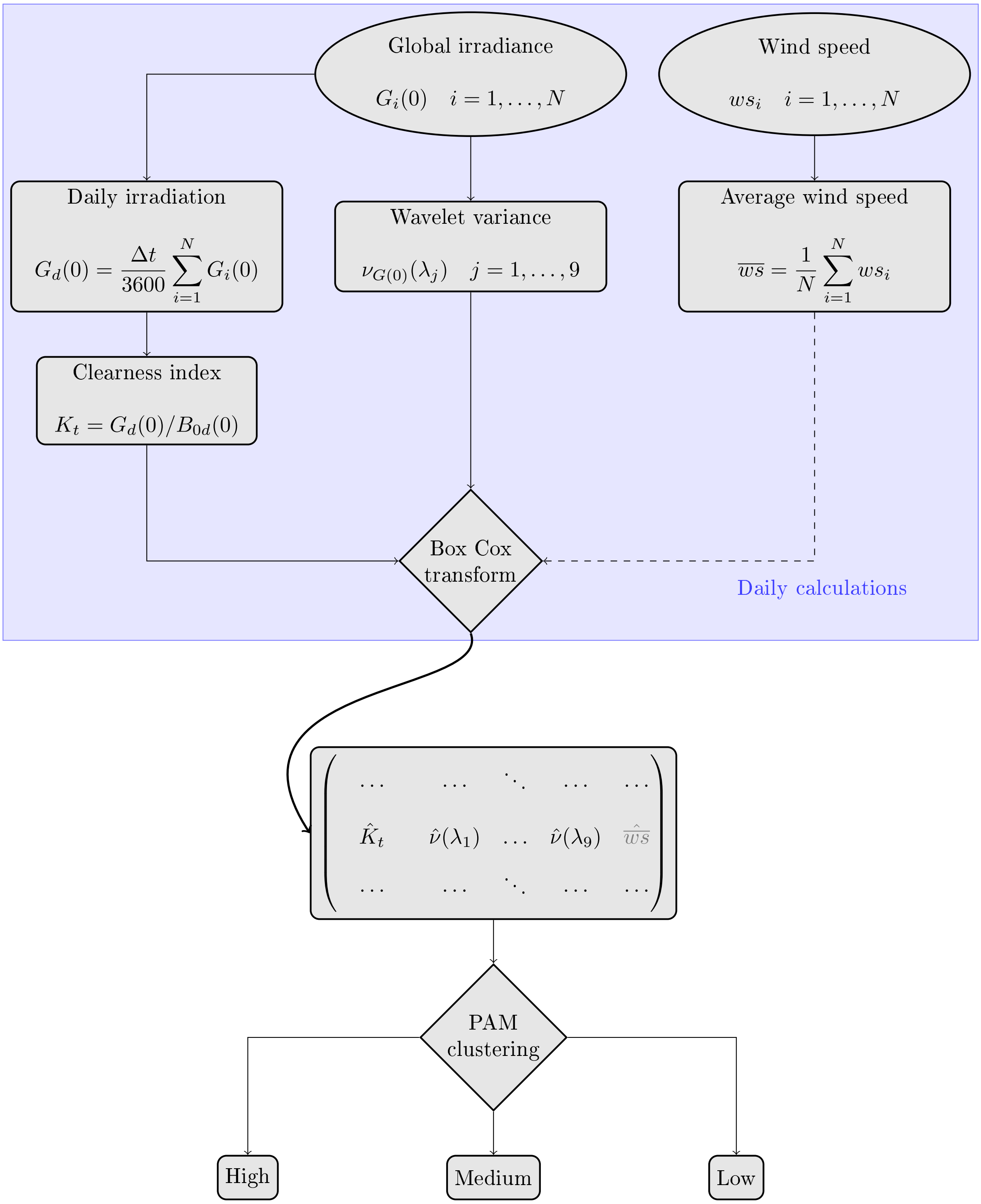


Figure 2: Feature generation and clustering algorithm. The blue frame enclose the set of calculations performed on a daily basis. The result of this frame fills a row in the matrix constructed with the global dataset. The dashed line and gray color of the average wind speed denote its poor contribution to the PAM clustering results.

the pairwise correlation between inverters at different scales for each day of the data set (section 3.3).

We include here only a summary of the main steps of the wavelet analysis (figure 1). The interested reader is referred to the book [14].

3.2.1. *Multiresolution analysis*

Let $\mathbf{X}$ be an N dimensional vector containing the real-valued time series $X_t : t = 0, \ldots, N-1$, where the sample size N is any positive integer. For any positive integer J_0, the level J_0 of the Maximum Overlap Discrete Wavelet Transform (MODWT)[3] of $\mathbf{X}$ is a transform, $\widetilde{\mathbf{W}}_\mathbf{j} = \widetilde{\mathcal{W}}_j \mathbf{X}$, which produces a set of $J_0 + 1$ N-dimensional vectors, $\widetilde{\mathbf{W}}_\mathbf{1} \ldots \widetilde{\mathbf{W}}_{\mathbf{J_0}}$ and $\widetilde{\mathbf{V}}_{J_0}$ [14]. $\widetilde{\mathbf{W}}_\mathbf{j}$ is the vector of MODWT wavelet coefficients related with changes on scale $\lambda_j \equiv 2^{j-1}$, and $\widetilde{\mathbf{V}}_{\mathbf{J_0}}$ is the vector of MODWT scaling coefficients associated with averages on scale 2^{J_0}. The number of decomposition levels, J_0, is limited by the length of the signal through $N = 2^{J_0}$.

The elements of $\widetilde{\mathbf{W}}_\mathbf{j}$ and $\widetilde{\mathbf{V}}_{\mathbf{J_0}}$ are the output of a linear filtering operation implemented with MODWT wavelet and scaling filters. $\widetilde{\mathcal{W}}_j$ and $\widetilde{\mathcal{V}}_j$ are the mathematical representation of these filters, named the MODWT wavelet and scaling matrices, respectively. Among the variety of available filters the calculations reported in this paper have been performed with the Daubechies least asymmetric (*symmlet*)[4] filters with filter length $L = 8$ [14].

The time series $\mathbf{X}$ can be recovered from the MODWT with a multiresolution analysis (MRA)[5]:

$$\mathbf{X} = \widetilde{\mathcal{W}}^T \widetilde{\mathbf{W}} = \sum_{j=1}^{J_0} \widetilde{\mathcal{W}}_j^T \widetilde{\mathbf{W}}_j + \widetilde{\mathcal{V}}_{J_0}^T \widetilde{\mathbf{V}}_{J_0} \equiv \sum_{j=1}^{J_0} \widetilde{\mathcal{D}}_j + \widetilde{\mathcal{S}}_{J_0} \tag{1}$$

where $\widetilde{\mathcal{D}}_j = \widetilde{\mathcal{W}}_j^T \widetilde{\mathbf{W}}_j$ is the j-th level detail and $\widetilde{\mathcal{S}}_{J_0} = \widetilde{\mathcal{V}}_{J_0}^T \widetilde{\mathbf{V}}_{J_0}$ is the J_0-th level smooth.

3.2.2. *Wavelet variance*

The time-dependent wavelet variance for scale λ_j is defined as the variance of the wavelet coefficients at level j[6]:

$$\nu^2_{X,j} = \frac{1}{N} \sum_{t=0}^{N-1} \widetilde{W}^2_{j,t} = \frac{1}{N} |\widetilde{\mathbf{W}}_\mathbf{j}|^2 \tag{2}$$

The energy decomposition of the MODWT can be combined with this definition to show that the wavelet variance decomposes the variance of certain stochastic processes on a scale basis. The energy decomposition of $\mathbf{X}$ is:

$$|\mathbf{X}|^2 = |\widetilde{\mathbf{W}}|^2 = \sum_{j=1}^{J_0} |\widetilde{\mathbf{W}}_j|^2 + |\widetilde{\mathbf{V}}_{J_0}|^2 \tag{3}$$

Since the variance of the time series X_t is $\sigma^2_X = \frac{1}{N}|\mathbf{X}|^2 - \overline{X}^2$, and $\frac{1}{N}|\widetilde{\mathbf{V}}_{\mathbf{J_0}}|^2 \simeq \overline{X}^2$, equations (2) and (3) yield:

[3] The function `wavMODWT` of the package `wmtsa` performs the MODWT of a time series.

[4] The function `wavDaubechies` of the package `wmtsa` computes this filter using `wavelet = 's8'`.

[5] The function `wavMRD` of the package `wmtsa` calculates the detail from a MODWT and the function `reconstruct` can invert the wavelet transform to the original series.

[6] For ease of exposition, the equation includes the whole set of coefficients. However, it is recommended the use of the unbiased wavelet variance. This estimator avoids those coefficients subject to circular filter operations (boundary coefficients)

Wavelet scale	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
Time scale	10 to 20 s	20 to 40 s	40 to 80 s	1.33 to 2.67 min	2.67 to 5.33 min	5.33 to 10.67 min	10.67 to 21.33 min	21.33 to 42.67 min	42.67 to 85.33 min	1.42 to 2.84 h

Table 1: Physical time scales corresponding to each wavelet scale.

$$\sigma_X^2 = \frac{1}{N}\sum_{j=1}^{J_0}|\widetilde{\mathbf{W_j}}|^2 \tag{4}$$

Therefore, the wavelet variance decomposes the variance of the time series $\mathbf{X}$[7]:

$$\sigma_X^2 = \sum_{j=1}^{J_0}\nu_{X,j}^2 \tag{5}$$

Because the wavelet variance $\nu_X^2(\lambda_j)$ is just the variance of the MODWT wavelet coefficients at scale λ_j, the relationship between wavelet scale and frequency leads to an alternative summary of the spectral density function (SDF) [14, 16]:

$$\nu_{X,j}^2 = \frac{\overline{S_{X,j}}}{2^j\Delta t} \tag{6}$$

Equation (6) shows that the SDF is summarized with the wavelet variance using the average value per octave frequency band:

$$\overline{S_{X,j}} = 2^{j+1}\Delta t\int_{\frac{1}{2^{j+1}\Delta t}}^{\frac{1}{2^j\Delta t}} S_X(f)\mathrm{df} \tag{7}$$

The width of the octave of the correspondent scale λ_j is $1/(2^{j+1}\Delta t)$, where Δt is the sampling time of the signal. The frequency band of this scale is $1/(2^{j+1}\Delta t) \leq f \leq 1/(2^j\Delta t)$ (Table 1).

3.3. *Cross-correlation of wavelet time series*

The ability of the MODWT to capture variability in both time and scale can be extended to show the bivariate relationship between two time series both locally in time and frequency.

The wavelet cross-covariance of two stochastic processes X_t and Y_t for scale $\lambda_j = 2^{j-1}$ and lag τ is defined as [24]:

$$\gamma_{\tau,XY}(\lambda_j) \equiv \mathrm{Cov}\{\overline{W}_{j,t}^X, \overline{W}_{j,t+\tau}^Y\} \tag{8}$$

where $\overline{W}_{j,t}^X$ and $\overline{W}_{j,t}^Y$ are the λ_j MODWT coefficients of each process. These coefficients have mean zero for adequate wavelet filters, and therefore:

$$\gamma_{\tau,XY}(\lambda_j) = \mathrm{E}\{\overline{W}_{j,t}^X, \overline{W}_{j,t+\tau}^Y\} \tag{9}$$

[7] The function `wavVar` of the package `wmtsa` estimates the wavelet variances from a time series.

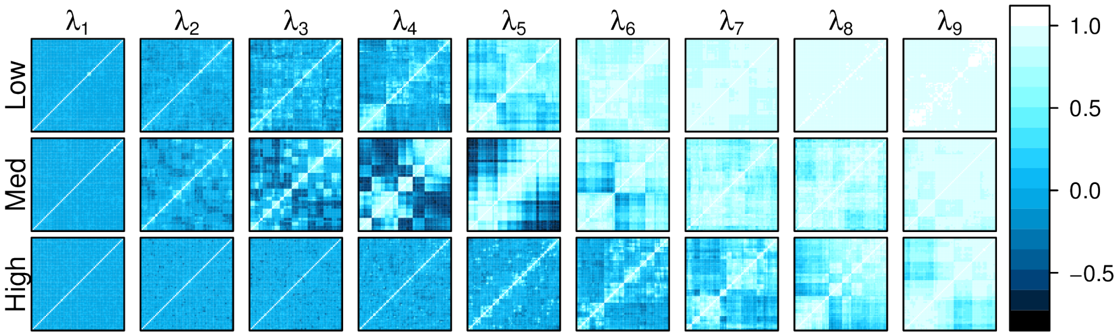


Figure 3: Correlogram matrix with the three typical days and the wavelet scales.

Similarly the wavelet cross-correlation for scale λ_j and lag τ can be defined with the wavelet cross-covariance and the wavelet variances for both processes[8]:

$$\rho_{\tau,XY}(\lambda_j) \equiv \frac{\gamma_{\tau,XY}(\lambda_j)}{\nu_X(\lambda_j)\nu_Y(\lambda_j)} \tag{10}$$

Since this is just a correlation between two random variables, $-1 \leq \rho_{\tau,XY}(\lambda_j) \leq 1$ for all scales and lags. When the lag is zero ($\tau = 0$) we obtain the wavelet covariance and correlation, which will be denoted as $\gamma_{XY}(\lambda_j)$ and $\rho_{XY}(\lambda_j)$ respectively to simplify notation.

Here we use the wavelet correlation to analyse the scale-dependent correlations for the scales 1 to 9 between the power time series of each of the possible combinations between the 70 inverters of the plant. This results in a set of nine wavelet correlation matrices with 70 rows and columns for each day of the data set.

The figure 3 displays three correlograms for three typical days (explained in section 3.4.4 and displayed in figure 9) to explore the inner structure of the field of inverters. The figure 4 plots the correlation against the distances between inverters to analyse the behaviour of the wavelet correlation with distance for each scale.

3.3.1. *Uncertainty associated with estimators and experimental data*

Two important facts must be highlighted both directly related to using experimental data instead of simulations:

- Since the data is a partial realization of the stochastic process, there is a confidence interval associated with the MODWT estimators of the wavelet correlation [24]. The figure 5 displays this confidence interval for each scale and correlation value. It is wider for last wavelet scales and for low correlation values. Fortunately, our results show high correlation values for these scales (figure 3) and, therefore, the confidence intervals are narrow enough. Anyway, the confidence interval limits the comparison between points and must be considered throughout the analysis.

- Electrical power time series include the statistical differences between the performance of the inverters and PV generators. These differences are independent from the variability

[8] The function `cor` of the `stats` package produces for each wavelet scale a matrix of correlations between the correspondent wavelet coefficients calculated with the function `wavMODWT` of the `wmtsa` package.

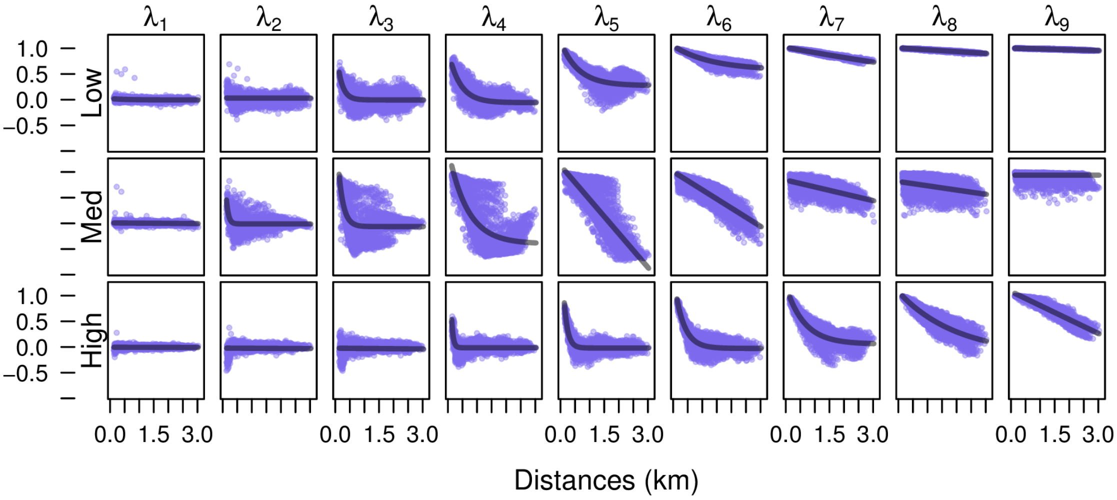


Figure 4: Wavelet correlation versus the distance between inverters. The lines correspond to the exponential decay model as defined in equation (14) and table 2.

of the irradiance and, although do not represent the key role of the relation between time series, they may reduce the correlation between time series.

3.4. Clustering

In order to relate the wavelet correlation with the fluctuation level, the data set is partitioned in three different groups according to the daily fluctuation behaviour of the global irradiance with the meteorological measurements registered by the PV plant weather station. This section details a clustering algorithm and a variable selection procedure to divide the matrix of daily in three different clusters of low, medium and high fluctuation levels (figure 2).

3.4.1. Partitioning Around Medoids

The unsupervised classification or clustering of data is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar to each other than to those in other clusters of the data. There is a wide variety of clustering techniques [17]. Due to its simplicity and robustness we have chosen the Partitioning Around Medoids (PAM) algorithm, which searches for k representative objects, called *medoids*, among the objects of the data set [18]. These *medoids* are computed such that the total dissimilarity of all objects to their nearest medoid is minimal. Therefore, the goal is to find a subset of k objects $m_1, \dots, m_k$ which minimizes the objective function:

$$\sum_{i=1}^{n} \min_{t=1,\dots,k} d(i, m_t) \qquad (11)$$

where d is a dissimilarity measure. The results reported in this paper are calculated with the euclidean distance.

After choosing the medoids, an object i is assigned into cluster t when the medoid m_t is nearer to i than the rest of medoids[9]:

[9] The `pam` function of the package `cluster` implements the PAM method.

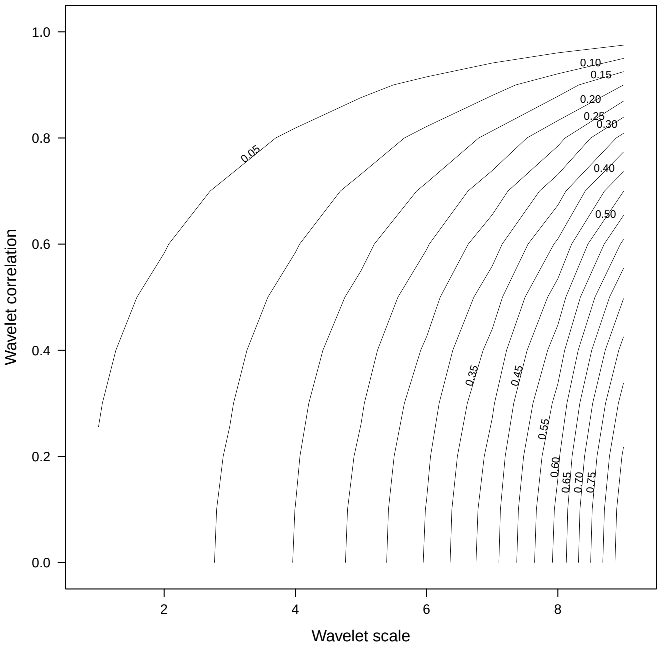


Figure 5: Confidence intervals associated with the MODWT estimator of the wavelet correlation.

$$d(i, m_t) \leq d(i, m_j) \quad \text{for} \quad j = 1, \ldots, k. \tag{12}$$

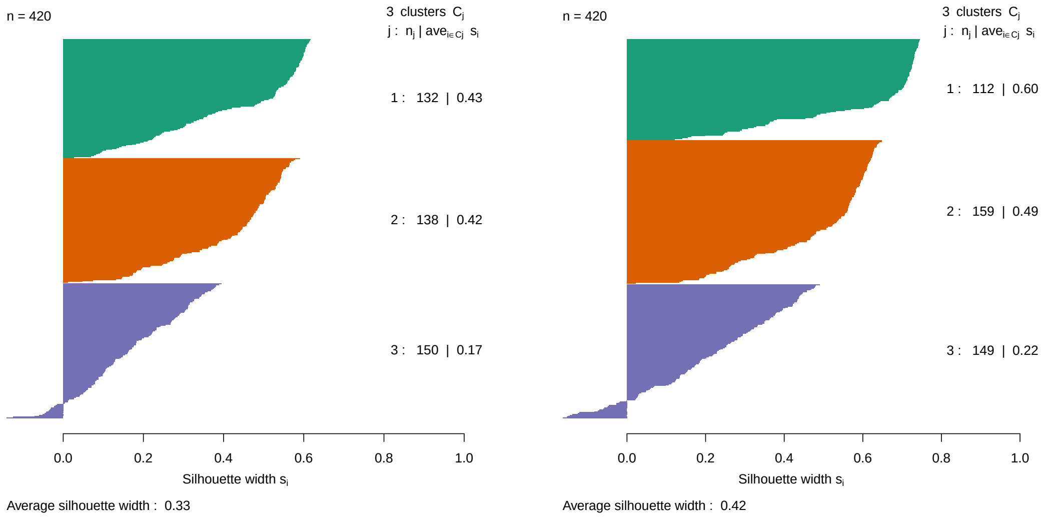
3.4.2. Variables selection

Before clustering, the data must be represented with a set of features (pattern representation). The goal of this feature generation is to discover compact and informative representations of the data. Since our subsequent interest is to break the collection into useful groups, the features should lead to large between-class distance and small within-class variance in the feature vector space. This means that features should take distant values in the different classes and closely located values in the same class [19].

In a clustering context with no class labels for patterns, the feature selection method involves a trial-and-error process where various subsets of features are selected, the resulting patterns clustered, and the output evaluated using a validity index [17].

In a previous paper Perpiñán and Lorenzo showed the behaviour of the solar irradiance as a long memory process and suggested the use of the exponent of the wavelet variance as a useful indicator of the fluctuation level [8]. In fact, this approach is one of the methods to estimate the fractal index of a time series, which is a measure of roughness (or smoothness) [20]. The fractal dimension (directly related to the fractal index) and the clearness index were proposed in [21] to classify solar irradiance in three different classes.

Although the fractal index (dimension) is a useful indicator of the fluctuation level as a summary of the information contained in the wavelet variances, its associated uncertainty [20] discourages its use in a clustering context. Thus, instead of using a derived variable (fractal dimension) we have decided to work with the primary variables (wavelet variances).



(a) With mean and maximum daily wind speed.

(b) Without mean and maximum daily wind speed.

Figure 6: Silhouette plots of the clustering results with (6a) and without (6b) wind speed values. In these graphics, "Low", "High" and "Medium" clusters (section 3.4.4) correspond to $j = 1$, $j = 2$ and $j = 3$, respectively.

For the clustering of *daily* data we have tested a set of features composed of the daily global irradiation, the clearness index[10], the wavelet variances and the mean and maximum daily wind speed. These features are tested with the PAM method and the results validated with the silhouette plot [22], a graphical display of the average dissimilarity between objects and medoids[11], whose average width is a suitable quality index to validate the clustering performance [18]. The tests results indicate that the wind speeds contribute negatively to the performance of the clustering procedure. Besides, the inclusion of the global irradiation distorts the clustering results if the feature matrix is not previously scaled.

Figure 6 compares the silhouette plots of the clustering results with and without mean and maximum daily wind speed (global irradiation is excluded from both plots). The average silhouette width with wind speeds is 0.37 and increases to 0.48 if these features are not included in the matrix. It is interesting to note that in both graphics clusters 1 and 2 have better silhouette widths than cluster 3. As it will be shown in section 3.4.4, clusters 1 and 2 contains days with low and high levels of power fluctuation, respectively, and cluster 3 contains days with medium levels of fluctuation.

It is important to stress that, although the relation between wind speed and irradiance fluctuation has been previously suggested [8, 13], our experimental analysis does not provide evidences to support this relation. Instead, other authors connect the irradiance changes with the cloud transit speed [4] which cannot be directly measured but estimated with the cloud motion analysis from satellite images.

[10] The `fSolD` function of the `solaR` package calculates the daily extraterrestial irradiation.

[11] The `silhouette` function of the `cluster` package calculates the silhouette of a PAM clustering.

Consequently, the final set of features only comprises the clearness index and the wavelet variances.

3.4.3. Box-Cox transformation

This set of features provides a matrix of values whose distributions functions are strongly positively skewed. Before using this matrix with the clustering algorithm a transformation is recommended [19]. The family of the Box-Cox power functions [23] create a rank-preserving transformation and are recognised as a useful data pre-processing technique used to stabilise variance and make the data more normal distribution-like.

Let x be the original feature and x_ϵ the transformed feature. Then:

$$x_\epsilon = \begin{cases} \frac{x^\epsilon - 1}{\epsilon} & \text{if } \epsilon \neq 0 \\ \log(x) & \text{if } \epsilon = 0 \end{cases} \tag{13}$$

where ϵ is calculated for each feature with the Box-Cox method[12] [23].

3.4.4. Clustering results

The result of the PAM clustering method applied to the transformed matrix of features is displayed in the figures 7 and 8.

The figure 7 shows a scatter plot matrix where all the variables are confronted together (including the global irradiation and wind speeds) with their kernel density estimations in the diagonal frames. Different colors are asignated to each cluster, labelled as "High", "Medium" and "Low" according to their fluctuation levels as represented by the wavelet variances. The linear relation between the wavelet variances is easily appreciable. The clearness index and the wavelet variances are connected with a non-linear relation. The days with low fluctuations levels have very high (clear days) or very low (overcast days) values of clearness index, while days with middle and high fluctuation levels can be found in the middle range of the clearness index. Finally, the wind speed values do not show a precise relation neither with the clearness index nor with the wavelet variances.

The figure 8 displays the kernel density estimates of the wavelet variances grouped by clusters (the same colours of the figure 7 are being used here). It is evident that each cluster is clearly separated for these variables.

This procedure has been applied to the whole data set. For ease of exposition, the results will be illustrated in section 4 using only three days extracted from each cluster (figure 9). It must be underlined that there are nonnegligible differences among the collection of days belonging to a same cluster (figure 6b and section 3.3.1). However, although the results exposed with these three days is not exactly repeated by the rest of days of the corresponding cluster, we have found that the wavelet correlation follows a common pattern for the days of a cluster. This common structure will be highlighted further and, therefore, the analysis and conclusions to be detailed are valid for almost every day of each cluster.

The figure 10 displays the maximum irradiance fluctuation at each wavelet scale for these three days. The maximum irradiance fluctuation for the scale λ_j is estimated with the maximum value of the j-th level wavelet coefficients vector, $\widetilde{\mathbf{W}}_j$, normalized with the STC irradiance, $G_{STC} = 1000\,\frac{\mathrm{W}}{\mathrm{m}^2}$.

[12] The functions `powerTransform` and `bcPower` of the package `car` implement the Box-Cox family functions.

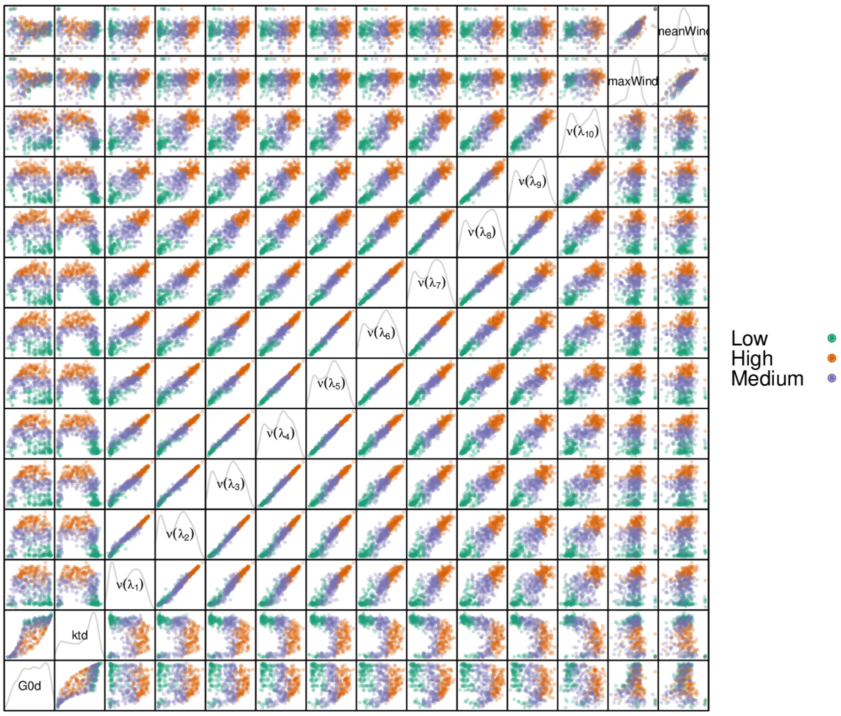


Figure 7: Scatterplot matrix of features tested with the quality index of the PAM algorithm. Colors indicate the cluster membership as determined by the PAM method.

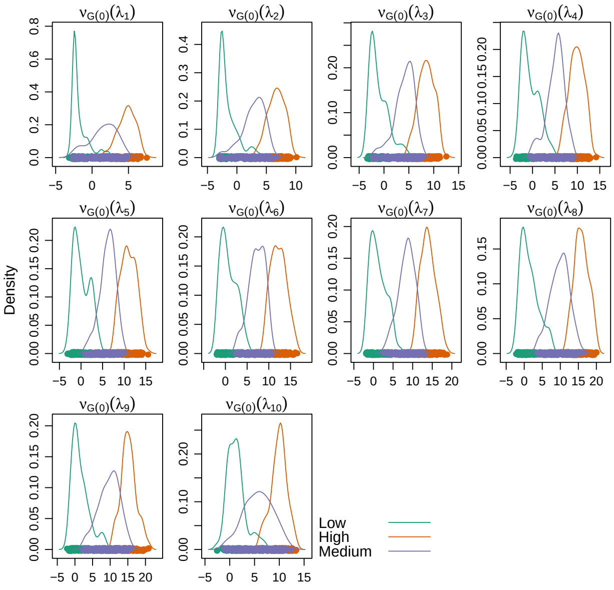


Figure 8: Kernel density estimates of the wavelet variances grouped by clusters. The x-scales displays the wavelet variances after the Box-Cox transformations.

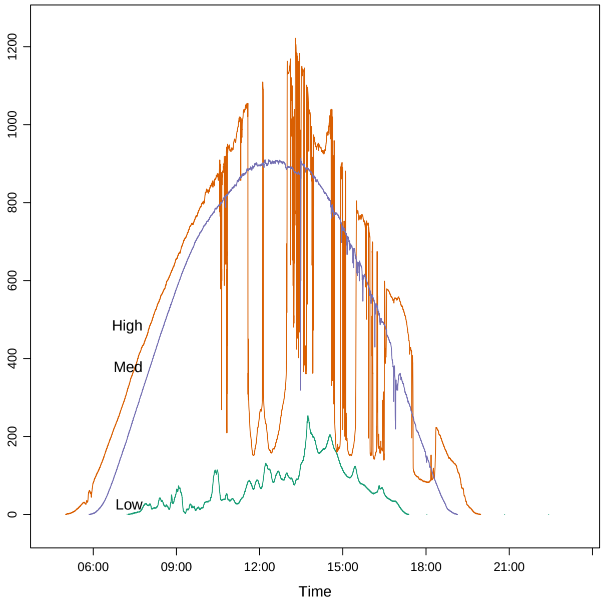


Figure 9: Typical days from each cluster labelled according to the fluctuation level.

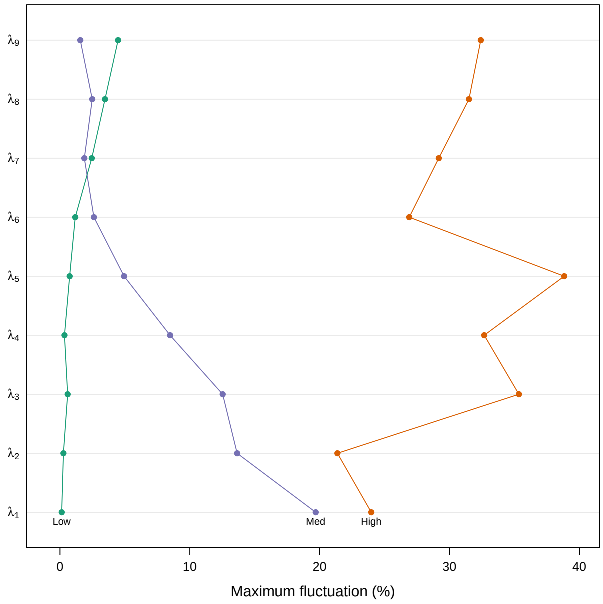


Figure 10: Maximum irradiance fluctuation at each wavelet scale for the three fluctuation levels.

4. Discussion

The correlograms of the figure 3 show low correlation values for the two lowest wavelet scales for the three classes. This behaviour does not change for the wavelet scales λ_3 and λ_4[13] with data from the class "High". However, the correlation matrices of these wavelet scales in classes "Low" and "Medium" show an increasing pattern: several pairs of inverters are fluctuating in synchronism (positive values) or in opposition (negative values). In the "Low" cluster the positive values of correlation are predominant, while in the "Medium" cluster the correlation values range from $+1$ to -0.5. The wavelet scales λ_5 and λ_6 in class "Low" are dominated by high positive values. The class "Medium" at these scales still shows patterns of high positive and high negative values, although the scale λ_6 is almost diluted with positive values. The class "High" at these scales starts to show a weak pattern with positive values which is increasingly defined at scales λ_7, λ_8 and λ_9. These wavelet scales of classes "Low" and "Medium" are defined almost completely with values next to 1.

The figure 4 displays the same behaviour confronting the correlation with the distance between inverters. Low correlation values for the whole range of distances is evident for the scale λ_1 for the three classes. The wavelet scales λ_3 and λ_4 of the class "High" still show low correlation values independently of the distance. The correlation at these wavelet scales in classes "Low" and "Medium" are high for short distances decaying rapidly with distance. In the "Medium" cluster the correlation is positive for short distances and changes to negative values with increasing distances. This behaviour is particularly visible at λ_5 and λ_6 scales. On the other hand, in the "Low" cluster the correlation is almost always positive. Particularly, the wavelet scales λ_6 to λ_9 in this class are dominated by increasingly high positive values. The class "Medium" at the scales λ_7 to λ_9 also shows a flatter behaviour with the distance although with a higher dispersion. The class "High" is dominated by low correlation values and a flat relation with the distance through scales λ_1 to λ_4. At scale λ_5 the correlation gets higher at short distances with a fast decaying response. The scales λ_6 and λ_7 reinforce this behaviour resulting in a quasi-linear relation at scales λ_8 and λ_9 with correlation values next to 1 for short distances reaching 0 at the largest distances.

In summary:

- The correlation values at the scales λ_1 and λ_2 (corresponding to the time periods 10 to 20 s and 20 to 40 s respectively) are low without dependence on the fluctuation level of the day.

- The correlation at the scales λ_8 and λ_9 (corresponding to the time periods 21.33 to 42.67 min and 42.67 to 85.33 min respectively) are positive and high values. With low fluctuation levels the correlation is close to one for the whole range of distances and for every combination of inverters. When the fluctuation level is medium the correlation ranges from 0.5 to 1 slowly decaying with the distance. With high fluctuation levels the correlation values extend from 0 to 1 decreasing strongly with distance.

- The correlation behaviour at the intermediate scales depends on the fluctuation level. For example, the scales λ_3 and λ_4 (40 to 80 s and 1.33 to 2.67 min with low fluctuation levels are similar to the scales λ_5 and λ_6 (2.67 to 5.33 min and 5.33 to 10.67 min) with high fluctuation levels.

[13] The table 1 relates the wavelet scale with the physical time scale.

4.1. Relation with previous research

This tendency is consistent with recent investigations. In particular, the references [8, 25] propose to model a PV plant as a low-pass filter: high frequencies (scales λ_1 and λ_2) are strongly attenuated because the correlation values are very low; low frequencies remain approximately unchanged since the correlation values are close to 1; intermediate frequencies filtering depends on the distance and on the fluctuation level of the day.

Moreover, these results show agreement with the "wavestrapping" method [26] proposed in [8]. This method produces new versions of an original irradiance signal with the same fluctuation behaviour. The diurnal trend of the original time series is unchanged while the detrended irradiance is wavestrapped. The wavelet coefficients of the detrended irradiance (the sum of the first wavelet scales) are assumed to be from an independent and identically distributed population so new time series can be constructed with random sampling with replacement from the original decomposition. This approach is consistent with the low levels of cross-correlation between different points of the plant at the first wavelet scales.

On the other hand, Hoff and Perez [4] developed a model to quantify the output variability from an ensemble of identical PV systems. Under this model, when the change in output between locations is uncorrelated for the considered time scale, the fleet output variability equals the output variability at any one location divided by the square root of the number of locations. This relationship is the result of the sum of uncorrelated random variables with identical variance, and is consistent with our results at the λ_1 and λ_2 scales. For other cases, the variability predicted by this model depends on the distance and time scale considered, as confirmed with our experimental data.

4.2. Exponential decay model

The behaviour displayed in figure 4 can be modelled with a exponential decay model:

$$\rho(d) = \mathrm{a} + \mathrm{b} \cdot \exp(-\frac{d}{\mathrm{c}}) \tag{14}$$

where d is the distance in meters between inverters, a is the asymptotic value for large distances, $a + b$ is the correlation value for short distances, and c is the range factor. For example, the estimated correlation at wavelet scale λ_5 between two inverters separated 300 m is $-0.015 + 5.5 \cdot \exp(-300/57) = 0.013$ for the high fluctuation level. The reader is bewared that the use of this model outside the distances range of the figure 4 can result in erroneous correlation values, that is $|\rho(d)| > 1$.

The range factor is the distance in which the difference of the model from the asymptote becomes shorter than $0.37 \cdot b/a$:

$$\frac{\rho(c) - \rho(\infty)}{\rho(\infty)} \simeq 0.37 \cdot \frac{b}{a} \tag{15}$$

$$\frac{\rho(3c) - \rho(\infty)}{\rho(\infty)} \simeq 0.05 \cdot \frac{b}{a} \tag{16}$$

Since this difference depends on the ratio between a and b, the range factors from two models with different coefficients a and b must be compared cautiously.

The table 2 contains the coefficients of this model for each of the combinations between wavelet scale and fluctuation level of the figure 4. Those combinations of wavelet scale and fluctuation level where the correlation is almost constant or linear with the distance can also be adjusted with this model. However, the corresponding coefficients are meaningless (the

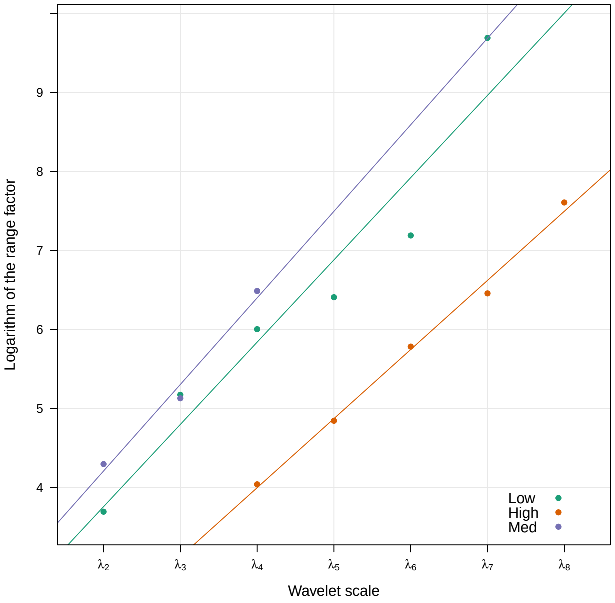


Figure 11: Logarithm of the range coefficient of equation (14) versus the wavelet scale for each fluctuation cluster.

absolute value of the asymptotic value of the correlation, $|a|$, is larger than 1, or the range factor, c, is very large or negative) and not useful for comparisons.

These combinations (marked in italic in the table 2) are excluded from the figure 11, which displays the logarithm of the range factor versus the wavelet scale for each fluctuation cluster. Loosely speaking, the range factor is higher for the "Medium" and "Low" clusters than for the "High" cluster, and it increases exponentially with the wavelet scale. In other words, the distance in which the correlation is similar to the value at large distances is shorter if the fluctuation level is high and for the first wavelet scales. Anyway, it must be underlined that the main purpose of the equation (14) is to better illustrate the relation between wavelet correlation, distance and wavelet scale and not to provide a model perfectly adjusted to the data. Therefore, the meaning of the model and its coefficients must be understood in this context.

5. Conclusions

This paper analyzes the correlation between the fluctuations observed in the power generated by the ensemble of 70 DC/AC inverters from a 45.6 MW PV plant, separated between 220 m to 2.8 km. The use of real electrical power time series from a large collection of photovoltaic inverters of a same plant is a major contribution in the context of models built upon simplified assumptions to overcome the absence of such data. Nevertheless, it must be underlined that there is an uncertainty associated to the use of experimental data: on the one hand, the confidence interval of the estimators of the wavelet correlation; on the other hand, the statistical differences between the inverters performance and between PV generators.

This data set has been divided into three different groups with a clustering procedure according to the daily fluctuation level of the global irradiance. The clustering procedure performs correctly with the information provided by the clearness index and the wavelet variances for

Fluctuation Level	Wavelet scale	a	b	$a+b$	c
Low	λ_1	*-0.005*	*0.026*	*0.021*	*5.7e+02*
	λ_2	0.031	0.032	0.063	40
	λ_3	-0.0051	1.1	1.127	1.8e+02
	λ_4	-0.058	1	0.973	4e+02
	λ_5	0.27	0.86	1.133	6.1e+02
	λ_6	0.57	0.48	1.049	1.3e+03
	λ_7	-0.68	1.7	1.019	1.6e+04
	λ_8	*3.4e+04*	*-3.4e+04*	*1.008*	*-9.5e+08*
	λ_9	*7.6e+03*	*-7.6e+03*	*1.005*	*-4.8e+08*
Medium	λ_1	*-8.3e+04*	*8.3e+04*	*0.009*	*1.8e+10*
	λ_2	-0.012	2.8	2.770	73
	λ_3	-0.064	2.2	2.152	1.7e+02
	λ_4	-0.4	1.9	1.461	6.6e+02
	λ_5	*-1e+06*	*1e+06*	*1.126*	*1.5e+09*
	λ_6	*1.3e+06*	*-1.3e+06*	*1.007*	*-3.6e+09*
	λ_7	*-5.3e+05*	*5.3e+05*	*0.843*	*3.9e+09*
	λ_8	*-8.5e+02*	*8.6e+02*	*0.814*	*1e+07*
	λ_9	*-7.3e+13*	*7.3e+13*	*0.938*	*1.2e+20*
High	λ_1	*-4.9e+04*	*4.9e+04*	*-0.005*	*3.3e+10*
	λ_2	*-1.6*	*1.6*	*-0.022*	*4.8e+05*
	λ_3	*-1.4*	*1.4*	*-0.023*	*2.7e+05*
	λ_4	-0.015	5.5	5.486	57
	λ_5	-0.02	2.5	2.448	1.3e+02
	λ_6	-0.027	1.4	1.415	3.2e+02
	λ_7	0.056	1.1	1.192	6.4e+02
	λ_8	-0.16	1.2	1.074	2e+03
	λ_9	*-7.3e+05*	*7.3e+05*	*1.081*	*2.7e+09*

Table 2: Coefficients of the exponential decay model expressed in equation (14) adjusted with the data displayed in figure 4. There are some combinations whose coefficients are meaningless (the absolute value of the asymptotic value of the correlation, $|a|$, is larger than 1, or c is very large or negative) and are marked in italic.

different time scales. Neither the global irradiation nor the wind speed contributes to improve the performance of the classification method.

Afterwards, the time dependent correlation between the electrical power time series of the inverters has been calculated using the wavelet transform. The results show that the wavelet correlation depends on the distance between the inverters, the wavelet time scales and the daily fluctuation level. Correlation values for time scales below one minute are low without dependence on the daily fluctuation level. However, for time scales above 20 minutes, positive high correlation values are obtained, and the decay rate with the distance depends on the daily fluctuation level. At intermediate time scales the correlation depends strongly on the daily fluctuation level: for example, 2 min fluctuations from a group of inverters may be uncorrelated if the day belongs to the “High” cluster, but show almost perfect correlation if the day falls in the “Medium” group.

These findings are consistent with recent investigations which model a PV plant as a low-pass filter: high frequencies (scales λ_1 and λ_2) are strongly attenuated because the correlation values are very low; low frequencies remain approximately unchanged since the correlation values are close to 1; intermediate frequencies filtering depends on the distance and on the daily fluctuation level.

The methods proposed in this paper have been implemented using the free software environment R [27]. The source code is available at https://github.com/oscarperpinan/wavCorPV.

Acknowledgements

The authors wish to thank Acciona Solar, who kindly allowed for the measurement of the PV plant under analysis.

References

[1] K. Otani, J. Minowa, K. Kurokawa, Study on areal solar irradiance for analyzing areally-totalized PV systems, Solar Energy Materials and Solar Cells 47 (1997) 281–288.

[2] O. K. Murata A, Yamaguchi H, A method of estimating the output fluctuation of many photovoltaic power generation systems dispersed in a wide area, Electrical Engineering in Japan 166 (4) (2009) 9–19.

[3] E. Wiemken, H. G. Beyer, W. Heydenreich, K. Kiefer, Power characteristics of PV ensembles: experiences from the combined power production of 100 grid connected PV systems distributed over the area of germany, Solar Energy 70 (6) (2001) 513–518.

[4] T. Hoff, R. Perez, Quantifying PV power output variability, Solar Energy 84 (10) (2010) 1782 – 1793. doi:DOI:10.1016/j.solener.2010.07.003.
URL http://www.asrc.cestm.albany.edu/perez/2010/short.pdf

[5] T. E. Hoff, R. Perez, Modeling PV fleet output variability, Solar Energy (0) (2011) –. doi:10.1016/j.solener.2011.11.005.

[6] J. Marcos, L. Marroyo, E. Lorenzo, D. Alvira, E. Izco, Power output fluctuations in large scale PV plants: One year observations with 1 second resolution and a derived analytic model., Progress in Photovoltaics.
URL http://138.4.46.62:8080/ies/ficheros/2_52_ref14.pdf

[7] J. Marcos, L. Marroyo, E. Lorenzo, M. García, Smoothing of PV power fluctuations by geographical dispersion, Progress in Photovoltaics: Research and Applications 20 (2) (2012) 226–237. doi:10.1002/pip.1127.
URL http://138.4.46.62:8080/ies/ficheros/2_52_ref15.pdf

[8] O. Perpiñán, E. Lorenzo, Analysis and synthesis of the variability of irradiance and PV power time series with the wavelet transform, Solar Energy 85 (1) (2011) 188 – 197. doi:DOI:10.1016/j.solener.2010.08.013.
URL http://procomun.wordpress.com/documentos/articulos/

[9] M. Lave, J. Kleissl, E. Arias-Castro, High-frequency irradiance fluctuations and geographic smoothing, Solar Energy 86 (8) (2012) 2190 – 2199. doi:10.1016/j.solener.2011.06.031.

[10] M. Súri, T. Huld, E. D. Dunlop, M. Albuisson, M. Lefevre, L. Wald, Uncertainties in photovoltaic electricity yield prediction from fluctuation of solar radiation, in: 22nd European Photovoltaic Solar Energy Conference, 2007.
URL http://re.jrc.ec.europa.eu/pvgis/doc/paper/2007-Milano_6DV.4.44_uncertainty.pdf

[11] O. Perpiñán, Statistical analysis of the performance and simulation of a two-axis tracking PV system, Solar Energy 83 (11) (2009) 2074–2085.
URL http://procomun.wordpress.com/documentos/articulos

[12] R. Perez, S. Kivalov, J. Schlemmer, K. H. Jr., T. E. Hoff, Short-term irradiance variability: Preliminary estimation of station pair correlation as a function of distance, Solar Energy 86 (8) (2012) 2170 – 2176. doi:10.1016/j.solener.2012.02.027.

[13] A. Woyte, R. Belmans, J. Nijs, Fluctuations in instantaneous clearness index: Analysis and statistics, Solar Energy 81 (2007) 195–206.

[14] D. Percival, A. T. Walden, Wavelet Methods for Time Series Analysis, Cambridge University Press, 2006.

[15] M. B. Priestley, Evolutionary spectra and non-stationary processes, Journal of the Royal Statistical Society. Series B (Methodological) 27 (1965) 204–237.

[16] D. P. Percival, On estimation of the wavelet variance, Biometrika 82 (3) (1995) 619–631.
URL http://staff.washington.edu/dbp/PDFFILES/wavevar.pdf

[17] A. K. Jain, M. N. Murty, P. J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (1999) 264–323. doi:http://doi.acm.org/10.1145/331499.331504.

[18] A. Struyf, M. Hubert, P. Rousseeuw, Clustering in an object-oriented environment, Journal of Statistical Software 1 (4) (1997) 1–30.
URL http://www.jstatsoft.org/v01/i04

[19] S. Theodoridis, K. Koutroumbas, Pattern Recognition, Fourth Edition, Academic Press, 2009.

[20] T. Gneiting, H. Sevcikova, D. B. Percival, Estimators of fractal dimension: Assessing the roughness of time series and spatial data, Tech. rep., University of Washington, Department of Statistics (2010).
URL www.stat.washington.edu/research/reports/2010/tr577.pdf

[21] S. Harrouni, A. Guessoum, A. Maafi, Classification of daily solar irradiation by fractional analysis of 10-min-means of solar irradiance, Theoretical and Applied Climatology 80 (2005) 27–36, 10.1007/s00704-004-0085-0.
URL http://www.environmental-expert.com/Files/6063/articles/5217/J7LL30K7WLHH6N9T.pdf

[22] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (0) (1987) 53 – 65. doi:10.1016/0377-0427(87)90125-7.

[23] G. E. P. Box, D. R. Cox, An analysis of transformations, Journal of the Royal Statistical Society. Series B (Methodological) 26 (2) (1964) pp. 211–252.
URL http://www.jstor.org/stable/2984418

[24] B. Whitcher, P. Guttorp, D. B. Percival, Wavelet analysis of covariance with application to atmospheric time series, Journal of Geophysical Research-atmospheres 105 (D11) (2000) 14941–14962.
URL http://staff.washington.edu/dbp/PDFFILES/wavcov.pdf

[25] J. Marcos, L. Marroyo, E. Lorenzo, D. Alvira, E. Izco, From irradiance to output power fluctuations: the PV plant as a low pass filter, Progress in Photovoltaics: Research and Applications 19 (5) (2011) 505–510. doi:10.1002/pip.1063.
URL http://138.4.46.62:8080/ies/ficheros/2_52_ref16.pdf

[26] D. Percival, S. Sardy, A. Davison, Nonlinear and nonstationary signal processing, Cambridge University Press, 2000, Ch. Wavestrapping Time Series: Adaptive Wavelet-Based Bootstrapping, pp. 442–71.
URL http://www.unige.ch/math/folks/sardy/Papers/wavestrap.pdf

[27] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0 (2012).
URL http://www.R-project.org

[28] A. Zeileis, G. Grothendieck, zoo: S3 infrastructure for regular and irregular time series, Journal of Statistical Software 14 (6) (2005) 1–27.
URL http://www.jstatsoft.org/v14/i06/

[29] W. Constantine, D. Percival, wmtsa: Insightful Wavelet Methods for Time Series Analysis, R package version 1.0-5 (2010).
URL http://CRAN.R-project.org/package=wmtsa

[30] O. Perpiñán, solaR: Solar radiation and photovoltaic systems with R, Journal of Statistical Software 50 (9) (2012) 1–32.
URL http://www.jstatsoft.org/v50/i09/

[31] E. J. Pebesma, R. S. Bivand, Classes and methods for spatial data in R, R News 5 (2) (2005) 9–13.
URL http://CRAN.R-project.org/doc/Rnews/

[32] D. Sarkar, F. Andrews, latticeExtra: Extra Graphical Utilities Based on Lattice (2011).
URL http://R-Forge.R-project.org/projects/latticeextra/

[33] D. Sarkar, Lattice: Multivariate Data Visualization with R, Springer, New York, 2008, iSBN 978-0-387-75968-5.
URL http://lmdvr.r-forge.r-project.org

[34] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, K. Hornik, cluster: Cluster Analysis Basics and Extensions, R package version 1.14.2 (2012).

[35] J. Fox, S. Weisberg, An R Companion to Applied Regression, 2nd Edition, Sage, Thousand Oaks CA, 2011.
URL http://socserv.socsci.mcmaster.ca/jfox/Books/Companion