SPRINGER BRIEFS IN ENERGY

Luigi Fortuna
Giuseppe Nunnari
Silvia Nunnari

Nonlinear Modeling of Solar Radiation and Wind Speed Time Series

Springer

SpringerBriefs in Energy

More information about this series at http://www.springer.com/series/8903

Luigi Fortuna · Giuseppe Nunnari
Silvia Nunnari

Nonlinear Modeling of Solar Radiation and Wind Speed Time Series

Springer

Luigi Fortuna
Dipartimento di Ingegneria Elettrica
Elettronica e Informatica
Università degli Studi di Catania
Catania
Italy

Giuseppe Nunnari
Dipartimento di Ingegneria Elettrica
Elettronica e Informatica
Università degli Studi di Catania
Catania
Italy

Silvia Nunnari
Dipartimento di Ingegneria Elettrica
Elettronica e Informatica
Università degli Studi di Catania
Catania
Italy

ISSN 2191-5520 ISSN 2191-5539 (electronic)
SpringerBriefs in Energy
ISBN 978-3-319-38763-5 ISBN 978-3-319-38764-2 (eBook)
DOI 10.1007/978-3-319-38764-2

Library of Congress Control Number: 2016938671

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

This book is dedicated to our readers

Preface

Low carbon has become a global issue, as testified at the recent World Climate Summit 2015, held in Paris. Nowadays, the only viable alternative to this crisis is to develop as much as possible the use of the so-called renewable energy, as the possibility of obtaining clean energy through fusion processes being remote. Several forms of alternative sources of energy are present in nature almost in unlimited quantities, referred to as *renewable*, because they are continuously regenerated. The main source for renewable energies is the Sun. From the Sun are naturally derived accumulations of water to produce hydroelectric power, wind for aeolic turbine generators, and photovoltaics plants to generate electric energy. Also, from the photosynthesis process it is possible to derive energy from biomass.

A challenging problem with integrating renewable energy based plants, such as solar and wind speed ones, into electric grid is that these plants are intermittent. Thus, predicting the weather variables is of great interest for applications. There are essentially two ways to address this issue. One is by using Numerical Weather Forecasting (NWF) models, which are reliable but quite complex and require real-time information, which is usually available from Meteorological Agencies only. Furthermore, it has been pointed out that NWF models have high errors in forecasting meteo variables at local-scale areas and without appropriate postprocessing are often inferior to machine learning approaches. The other kinds of methods are represented by the so-called statistical modeling approaches, which are based on the use of past data recorded at the site of interest. The latter kinds of methods, compared to the former ones, require less computational efforts, but are appropriate for short-time horizons only. This book is devoted to study statistical prediction models for solar radiation and wind speed time series and asses their performance in the range (1, 24) hours. Furthermore, the problem of classifying daily patterns of both solar radiation and wind speed will be addressed as a useful strategy to obtain statistical properties for longer prediction range. The book concisely describes the main techniques of time series analysis, with an emphasis on solar radiation and wind speed, since they are the main kinds of renewable energies involved in the production of electrical energy. The forecasting problem is

addressed by using the embedding phase space approach, which is one of the most powerful methods proposed in the literature for modeling complex systems. Further, the book will guide the reader in applying some machine learning techniques to classify the daily patterns; thus allowing to perform statistical analyses that are not possible by using traditional techniques. The concepts will be exposed as much as possible avoiding unnecessary mathematical details, focusing on very concrete examples in order to ensure a better understanding of the proposed techniques. Developing various topics, the readers will be guided on how to find the most appropriate software and data resources with which they could perform their own experiments. The structure of the book is as follows. Methods for analysis of time series are concisely reported in Chap. 1. Application of these methods to solar radiation and wind speed time series are described in Chaps. 2 and 3, respectively. Modeling approaches for solar radiation and wind speed time series, focusing essentially on nonlinear autoregressive (NAR) and Embedded Phase Space (EPS) models, are given in Chap. 4. Identification of solar radiation and wind speed hourly average prediction models is reported in Chaps. 5 and 6, respectively. Classification of daily patterns of solar radiation and wind speed time series are described in Chaps. 7 and 8, respectively. Concluding remarks are given in Chap. 9 and finally, a list of software functions and dataset mentioned in the book is included in Appendix A.

Catania
April 2016

Luigi Fortuna
Giuseppe Nunnari
Silvia Nunnari

Acknowledgments

The authors wish to thank Prof. Giorgio Guariso of the Politecnico di Milano for helpful discussions about the topics addressed in this book.

The authors also thank the University of Catania for the funding support under the grant FIR 2014.

Contents

Abbreviations

ADF	Augmented Dicky–Fuller test
ANFIS	Adaptive neuro-fuzzy inference systems
ARIMA	Autoregressive integrated moving average
ARMA	Autoregressive moving average
BP	Backpropagation
DFA	Detrended fluctuation analysis
EMD	Empirical mode decomposition
EPS	Embedding phase space
FF	Feedforward
FFNN	Feedforward neural networks
GARCH	Generalized autoregressive with conditional heteroskedasticity
GHI	Global horizontal irradiance
GPH	Geweke–Porter-Hudak Hurst exponent
MFDFA	Multifractal detrended fluctuation analysis
MISO	Multi-input single-output
MLP	Multilayer perceptron
NAR	Nonlinear autoregressive
NARX	Nonlinear autoregressive with exogenous inputs
NF	Neuro-fuzzy
NN	Neural network
NREL	National Renewable Energy Laboratory
pdf	Probability distribution density
PP	Phillips–Pearson test
PRWM	Persistent random walk model
R/S	Rescaled range analysis
TDNN	Time delay neural networks
VR	Variance ratio test
WWR	Western Wind Resource

Chapter 1
Time Series Methods

Abstract Time series analysis consists of methods devoted to asses some structural properties of the model considered for a given application. The most popular analysis methods are aimed to assess basic properties such as the stationarity, the embedding state-space dimension, and the degree of stability. Further relevant aspects dealing with time series are discovering of hidden patterns and cluster analysis. The main aim of this section is to introduce time series methods to perform all these kinds of analysis.

1.1 Stationarity Analysis

In mathematics and statistics, a stationary process is a stochastic process whose joint probability distribution does not change when shifted in time. Consequently, parameters such as the mean and variance, if they are present, also do not change over time and do not follow any trends. Non-stationarity arises when the mechanism producing the data changes in time. Of course, a time series could not be nonstationary in the long term but behave as stationary at short. Thus, dealing with this problem the length of considered time interval can be relevant. A simple way to assess non-stationarity of a time series is to compute the mean of the first half ($1 \leq i \leq N/2$) and second half ($N/2 + 1 \leq i \leq N$), where N is the number of points in the time series. The mean (or average) value of a time series $\{x_i\}$ can be computed as

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{1.1}$$

If the means of the two halves differ by more than a few standard errors for each half, then stationarity is almost certainly a problem. The standard error can be computed as $\frac{\sigma}{\sqrt{N}}$ where σ is the standard deviation which is expressed as in (1.2).

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2}. \tag{1.2}$$

L. Fortuna et al., *Nonlinear Modeling of Solar Radiation and Wind Speed Time Series*, SpringerBriefs in Energy, DOI 10.1007/978-3-319-38764-2_1

A time series whose first two moments (mean and variance) are constant is said to exhibit *weak stationarity*, but such a condition is generally insufficient for complex systems. Higher moments of the data can be examined, but they tend to be even less robust, and their time independence still does not guarantee stationarity for a chaotic system. Since most of the techniques available to model time series implicitly assume that the time series is stationary, at least in the considered time interval where data are available, it can be useful to know if this assumption can be considered true.

Several approaches are available to test for non-stationarity, such as the Dicky–Fuller test (*adftest*), the Phillips-Perron test (*pptest*) and the Kwiatkowski-Phillips-Schmidt-Shin test (*kpsstest*) which are based on assessing if a time series have a unitary root, or the variance ratio test (*vratiotest*) which is based on assessing if a time series is a random walk. The names given in parenthesis refer to the functions listed in A.1 suggested in this book to perform the mentioned tests.

1.2 Recurrence Plots

Recurrence plots are a useful tool to identify structures in a data set such as intermittency, which can be detected also by direct inspection, the temporary vicinity of a chaotic trajectory to an unstable periodic orbit, or non-stationarity. It essentially consists in scanning the time series and marking each pair of time indices (i, j) with a black dot, whose corresponding pair of delay vectors has distance $< \varepsilon$. In an ergodic situation, the dots should cover the (i, j)-plane uniformly on average, whereas non-stationarity expresses itself by an overall tendency of the dots to be close to the diagonal. For the purpose of stationary testing, the recurrence plots are not particularly sensitive to the choice of embedding. Recurrent plots shown can be computed by using the *recurr* function which is part of the TISEAN project [1].

1.3 Linear Detrending

Removing a linear trend from data allows to focus analysis on the fluctuations around the trend. A linear trend typically indicates a systematic increase or decrease in the data. A systematic shift can result, for example, from sensor drift. While trends can be meaningful, some types of analysis yield better insight once trends are removed. Furthermore, the effects of non-stationarity can often be reduced by detrending the data. Linear detrending in time series can be performed by using, for instance, the *detrend* function.

1.4 Noise Reduction

Filtering of signals from nonlinear systems requires the use of special methods since the usual spectral or others linear filters may interact unfavorably with the nonlinear structures. Nonlinear noise reduction does not rely on frequency information in order to define the distinction between signal and noise. Instead, structure in the reconstructed phase space will be exploited. The simplest nonlinear noise reduction algorithm replaces the central coordinate of each embedding vector by the local average of this coordinate. This simple approach is implemented as the *lazy* function in the TISEAN project.

1.5 Power Spectrum

A stationary or detrended time series can be represented by a superposition of sines and cosines of various amplitudes and frequencies. Indeed, the so-called Discrete Fourier Transform (DFT) states that the generic element x_i of a time series $\{x_i\}$ can be approximated as

$$x_i = \frac{a_0}{2} + \sum_{m=1}^{N} (a_m cos \frac{2\pi mi}{N} + b_m sin \frac{2\pi mi}{N}) \tag{1.3}$$

where N is the number of time series samples and (a_m, b_m) are pairs of coefficients which can be computed as

$$a_m = \frac{2}{N} \sum_{i=1}^{N} x_i cos \frac{2\pi mi}{N} \tag{1.4}$$

$$b_m = \frac{2}{N} \sum_{i=1}^{N} x_i sin \frac{2\pi mi}{N} \tag{1.5}$$

In other terms, using the DTF, the information hidden in x_i, is transformed in the frequency domain so that it is now represented by the N coefficients a_m, and b_m. The $N/2$ different frequencies (or armonics) $f = mf_0$ is limited on the low end to $f_0 = 1/N$ (the fundamental frequency). From DFT the so-called power spectral density, or more simply, the power spectrum can be computed. Since the power is proportional to the square of the amplitude of an oscillation and since there are both sine and cosine terms, the power $S(f)$ at frequency $f = mf_0$ is given by $S_m = a_m^2 + b_m^2$. There are various ways to normalize S_m, but since usually only relative values are of interest, this aspect is not relevant. The power spectrum density of a time series can be computed by using the *periodogram* function or the *spectrum* function.

1.6 Autocorrelation

The autocorrelation of a time series $\{x_i\}$, expressed by (1.6), measures how strongly on average each data point of a time series is correlated with the one k steps away.

$$G(k) = \frac{\sum_{i=1}^{N-k}(x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{N-k}(x_i - \bar{x})^2} \tag{1.6}$$

It is normalized in such a way that $G(0) = 1$. It can be shown that $G(k) = 0$ for uncorrelated data. Furthermore, it can be demonstrated that the autocorrelation function is the inverse Fourier transform of the power spectrum. The correlation function is only defined at integer values of k, but it can be considered a discrete sample of the continuous $G(t)$ that would result from correlating the continuous variable $x(t)$ from which the discrete time series x_i, was derived. In general, the correlation function falls from a value of 1 at $k = 0$ to zero at large k. The value of k at which it falls to $1/e \simeq 37\,\%$ is called the correlation time τ_c. For x_n nearly periodic, the correlation function will be a decaying oscillation, in which case τ_c is the time for the envelop to decay to $1/e$.

One of the problem dealing with the autocorrelation function is that it is a linear measure, each term of which (the lag-k autocorrelation coefficient) measures the extent to which x_n versus x_{n+k} is a straight line. This means that several nonlinear systems have nonlinear correlation. The correlation function is symmetric about $k = 0$, and so the full width is $2\tau_c$, which is a measure of how much *memory* the system has. The reciprocal of this quantity, $\frac{1}{2\tau_c}$, is an estimate of the average rate at which predictability of the time series is lost. The autocorrelation can be computed by using the MATLAB® *autocorr* function.

1.7 Mutual Information

As above-mentioned, a problem with the autocorrelation function is that it is a linear statistic and does not account for nonlinear correlations. To overcome this drawback and thus capture the nonlinear correlation it is possible to use the mutual information, defined as in (1.7)

$$I = -\sum_{i,j} p_{ij}(k) ln \frac{p_{ij}(k)}{p_i p_j} \tag{1.7}$$

where for some partition of the time series range, p_i is the probability to find a time series values in the ith interval and p_{ij} is the joint probability that an observation falls in the ith interval and the observation time k later falls into the jth interval. Mutual information is also useful for estimating the optimum embedding dimension in the embedding phase space approach that is considered in this book. The mutual information can be computed by using the *mutual* function.

1.8 Noise 1/f and Random Walks

The term $1/f$ noise is often used to denote the nature of the so-called long-term memory processes. The ubiquitous of $1/f$ noise is one of the oldest puzzles in contemporary physics. Long-memory processes have been observed in several fields such as physics, biology, astrophysics, geophysics, and sociology, just to mention a few. A heuristic argument indicates the special nature of $1/f$ fluctuations. Assume that $S(f) \sim 1/f^{\beta}$ and that $G(k) \sim 1/k^{\alpha}$. With relatively simple arguments it can be demonstrated that $1/f^{\beta} \sim 1/f^{1-\alpha}$. Thus when β is close to unity, the exponent α must be close to zero. For β exactly equal to unity then $G(k) \sim 1/k^{\alpha}$ becomes a a slow logarithmic decay. This means that power spectra of the form $S(f) \sim 1/f^{\beta}$ correspond to extremely long time correlation when $\beta \simeq 1$. Roughly speaking a $1/f^{\beta}$ noise has three main features:

- It has an autocorrelation that decays so slowly that its sum does not converge to a finite number. This means that the process is long-range dependent.
- The loglog power spectrum of a $1/f^{\beta}$ process is linear with slope β, where β usually ranges from $\beta = 0.5$ to $\beta = 1.5$. Note that a white noise, i.e., a completely uncorrelated time series, has a slope of $\beta = 0$, due to the fact that its energy is equally distributed for all frequencies and thus the power spectrum is flat. Instead, a random walk, i.e., a time series in which the differences between consecutive samples is a white noise, has $\beta = 2$.
- A third feature of a $1/f^{\beta}$ process is that it is self-similar, i.e., the statistical properties of the time series are the same regardless of the scale of measurement, and hence the process lacks a characteristic time scale.

1.9 Fractal Dimension and Hurst Exponent

One of the main aims of time series fractal analysis consists in computing its fractal dimension D and/or the Hurst exponent, using one of several methods proposed in literature.

1.9.1 The Box-Dimension

Among several definition of the fractal dimension, the most popular is probably the so-called box-dimension (referred to also as the capacity dimension), formally defined as

$$D = -\lim_{\varepsilon \to 0} \frac{log N_{\varepsilon}}{log \varepsilon} \tag{1.8}$$

In (1.8) ε is a small square lattice with side ε and N_ε is the number of grids needed to cover the time series.

Estimation of D is based on observing that if the relation between N_ε and ε is a power law of the type

$$N_\varepsilon \propto \varepsilon^{-D} \tag{1.9}$$

then by taking the log of both members in expression (1.9) we have

$$logN_\varepsilon = logC - Dlog\varepsilon \tag{1.10}$$

which represents a straight line in a log–log diagram, drawn in the plane N_ε versus ε. The angular coefficient of this line is D while C is a constant. Thus, in order to estimate D it will be enough to approximate the curve $logN_\varepsilon$ versus $log\varepsilon$ with a regression line by using the traditional least square approach. Several software tools are available to perform this calculation, such as the *boxcount* function, listed in Appendix A.1.

1.9.2 The Hurst Exponent

The Hurst exponent H is used to classify time series into types and gain some insight into their dynamics. Indeed, based on H, time series can be classified into three clusters, namely persistent, antipersistent, and uncorrelated time series. More precisely:

- $0.5 < H < 1$ indicates a time series with long-term positive autocorrelation, meaning that a high value in the series will probably be followed by another high value and that the values a long time into the future will also tend to be high.
- $0 < H < 0.5$ indicates a time series with long-term switching between high and low values in adjacent pairs, meaning that a single high value will probably be followed by a low value and that the value after will tend to be high, with this tendency to switch between high and low values lasting a long time into the future.
- $H = 0.5$ indicates a completely uncorrelated time series.

A variety of techniques exist for estimating the Hurst exponent, the most popular being the so-called R/S analysis [2], which can be summarized as follows.

For a given time series x_i, $i = 1, 2, \ldots, N$, consider its first $n \leq N$ values and compute the average value $\bar{x}$ and the standard deviation σ. The so-called $(R/S)_n$ statistic is defined as

$$(R/S)_n = \frac{1}{\sigma}[Max_n \sum_{i=1}^{n}(x_i - \bar{x}) - Min_n \sum_{i=1}^{n}(x_i - \bar{x})]. \tag{1.11}$$

One of the achievements of the R/S analysis is that

$$(R/S)_n \rightarrow Cn^H \tag{1.12}$$

as $n \rightarrow \infty$, being C a constant. Thus, by taking the log of both members in expression (1.12), we have

$$log(R/S)_n = logC + Hlog(n) \tag{1.13}$$

which suggests that, similarly to what described in Sect. 1.9.1, it is possible to estimate H by evaluating the slope of a straight line. Nevertheless, some authors [3] recognized that the original R/S analysis may show some drawback when the considered time series is not large enough and proposed some improvements. Referring to the list given in A.1, the R/S algorithm is implemented by the *hurst* function. Others approaches to compute the Hurst exponent, namely the DFA (detrended fluctuation analysis) analysis and the Geweke-Porter-Hudak (GPH) approach are implemented by the *dfa* and the *gph* functions, respectively.

Mandelbrot, the pioneer of fractal geometry, recognized that the Hurst exponent can be related to the fractal dimension of a time series by the simple expression $D = 2 - H$ [4]. However, such a theoretical expression it somewhat difficult to experimentally verify, since estimation of H and D depend on several factors such as the kinds of algorithms considered, the time series length, the sampling time, and so on. Furthermore, as clarified in the next section, time series often are not monofractal, as implicit assumed by the algorithms described in this section, but multifractal.

1.10 Multifractals

Most real fractal objects are not precisely self-similar and thus may have different dimension on different size scales and on different parts of the object. Fractals that can be fully characterized only by specifying a spectrum of dimensions are called multifractals. Roughly speaking, a multifractal can be considered as an interwoven set of fractals of different dimensions, each having a different weight. One of the techniques proposed in literature to perform the multifractal analysis of time series is the multifractal detrended fluctuation analysis (MFDFA) developed by [5] as an extension of the DFA, originally proposed by [6]. Such a kind of analysis has been successful applied in the field of biomedical signals by several authors (see [7] for a review of the subject). Furthermore, [7] implemented also a useful MATLAB® code, referred to as *MFDFA* in the list given in Appendix A, to perform the MFDFA analysis. One of the main achievements of the multifractal detrended fluctuation analysis is the so-called multifractal spectrum, also referred to as the singularity spectrum $f(\alpha)$, α being the so-called the Lipshitz-Holder exponent. Roughly speaking it is possible to say that while monofractal can be characterized by a single exponent (the fractal dimension), in order to characterize multifractal systems a continuous spectrum of exponents is needed. Multifractal systems are common in nature, especially

in geophysics. In Chaps. 2 and 3 it will be shown that solar radiation and wind speed time series belongs to the class of multifractal.

1.11 False Nearest Neighbors

A method to determine the minimal sufficient embedding dimension d is called the false nearest neighbor method. The idea is quite intuitive. Suppose the minimal embedding dimension for a given time series $\{x_i\}$ is d_0. This means that in a d_0 dimensional phase space the reconstructed trajectory is a one-to-one image of the trajectory in the original phase space. Thus the neighbors of a given point are mapped onto neighbors in the delay space. Due to the assumed smoothness of the dynamics, neighborhoods of the points are mapped onto neighborhoods again. This means that embedding in a d-dimensional space with $d < d_0$ the topological structure is no longer preserved and points are projected into neighborhoods of other points to which they would not belong in higher dimensions. These points are called false neighbors. The fraction of false nearest neighbors versus the embedding dimension can be estimated by using the *false_nearest* function listed in Appendix A.1.

1.12 Lyapunov Spectrum

The Lyapunov exponents are an important means of quantification of unstable systems, proposed in the framework of Chaos Theory. A bounded dynamical system with at least a positive Lyapunov exponent is chaotic, and the largest exponent describes the average rate at which predictability is lost. The Lyapunov exponents are close related to eigenvalues of a dynamical system, but there are important differences. For instance, while the Lyapunov exponents are always real numbers, the eigenvalues can be complex. However, both quantities can be determined from the Jacobian matrix assuming linear local dynamics. A system with n dimensions has n Lyapunov exponents.

In order to evaluate the predictability of a chaotic system, it can be enough to calculate the largest Lyapunov exponent only. The largest or the whole set of Lyapunov exponents, can be computed by using available software code such as, for instance, the *lyap_k* function and the *lyap_spec* function listed in Appendix A.1.

1.13 Daily Patterns

Usually geophysical time series, such as solar radiation, wind speed and air temperature, due to the Earth spinning, are affected by daily patterns. These patterns often are not clearly recognizable by a visual inspection of the measured time series

because of random phenomena affecting such series. An effective way to highlight the presence of the daily patterns is to eliminate these random phenomena by an appropriate averaging process. Formally, assuming that the daily patterns of the time series x_i are stored in a matrix $D(M, N)$ of appropriate dimensions, then a way to estimate the daily patterns DP hidden into x_i, is to compute expression (1.14).

$$\begin{aligned} DP(W_l, \tau) = & \frac{1}{W_l} \sum_{d=d_i}^{d=d_f} D(d, \tau), \\ & W_l = d_f - d_i + 1, d_f \geq d_i, \\ & \tau = 1, \ldots N, \\ & d = 1, 2, \ldots, n_y \end{aligned} \tag{1.14}$$

where:

- W_l is the window length, expressed in days, during which the averaging process is performed;
- d_i and d_f are the initial and final Julian day index which delimit the averaging window;
- τ is the time index within a day; and
- n_y is the number of days in a year.

An appropriate choice for the window length W_l could be 1 week or 1 month. For instance, assuming W_l equal to 1 month, the daily patterns of wind speed time series recorded at the station referred to as ID2257 are shown in Fig. 1.1.

As it is possible to observe, the amplitude of these patterns are accentuate in the summer months, when solar radiation is more intense, rather than in winter.

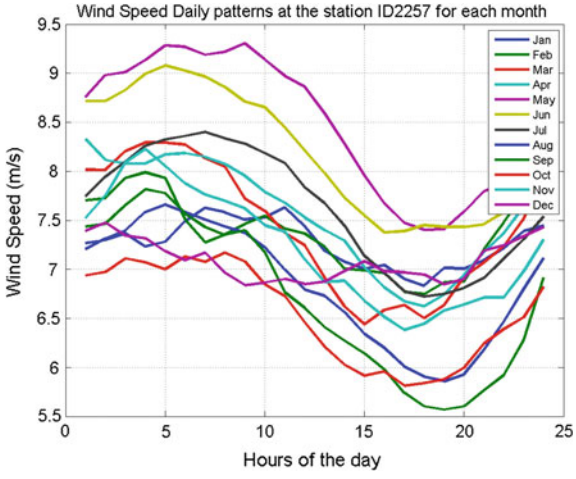

Fig. 1.1 Wind speed daily pattern at the ID2257 station

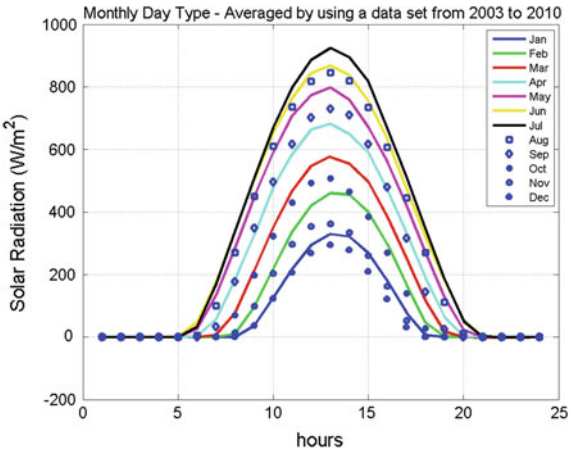

Fig. 1.2 An example of solar radiation typical day computed on monthly base at the Aberdeen station

Similarly, daily patterns of solar radiation time series, computed averaging on monthly basis the true time series recorded at the station referred to as Aberdeen, are shown in Fig. 1.2.

Also in this case, it can be observed that the averaging process eliminates random phenomena, giving to the daily patterns the typical bell-shaped. Of course, the amplitude and width of these bell curves is greater in summer than in winter time.

1.14 Time Series Clustering

Time series clustering can be useful to extract some information when others time series modeling approaches cannot provide reliable results. For instance, in our case, analysis performed on solar radiation and wind speed time series demonstrates (see the next Chaps. 2 and 3) that the average of daily values are scarcely autocorrelated, making rather difficult the prediction for the next day by using auto-regressive models. To partially overcome, this drawback, one may try to cluster the daily patterns (see Chaps. 7 and 8).

The goal of clustering is to identify possible structures in an unlabeled data set so that data is objectively organized in homogeneous groups. In other terms, clusters are formed by grouping objects that have maximum similarity with other objects within the group, and minimum similarity with objects in other groups. Clustering can be performed on static data, if all their feature values do not change with time, or change negligibly, or on time series.

Time series clustering is of interest due to the high pervasiveness of time series in various areas of science such as climatology, geology, business, and health sciences, just to mention a few [8]. A recent literature review of time series clustering can be found in [9].

Formally, the problem of time series clustering can be stated as follows: given a data set $D = \{x_1, x_2, \ldots x_n\}$ of n time series, partition D into $C = \{C_1, C_2, \ldots C_k\}$, such that $D = \bigcup_{i=1}^{k} C_i$. Each partition C_i represents a cluster containing at least one object. The partition is *crisp* if each object belongs to exactly one cluster, i.e., $C_i \bigcap C_j = \oslash$ for $i \neq j$, or *fuzzy* if one object is allowed to be in more than one cluster to a different degree of membership.

The heart of any clustering approach, regardless of whether the problem is to classify static data or time series is a clustering algorithm. Existing algorithms are of four kinds:

- exclusive clustering;
- overlapping clustering;
- hierarchical clustering; and
- probabilistic clustering.

1.14.1 The Exclusive Clustering

The k-means algorithm [10] is the prototype of exclusive clustering algorithms. It assigns to given set of patterns $\{x_k | k = 1, \ldots, n\}$ a predefined number k of cluster centers, expressed by a set of vectors V, $\{v_i | i = 1, \ldots c\}$, by minimizing an objective function of the form (1.15)

$$J(U, V) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \|x_k - v_i\|^2 \tag{1.15}$$

$u_{ik} \in \{0, 1\}, \forall i, k$ and $\sum_{i=1}^{c} u_{ik} = 1, \forall k$. Being u_{ik} a boolean expression, it is evident that the k-means requires a mutual exclusive belonging to the clustered patterns. The k-means algorithm is implemented in MATLAB® as the function *kmeans*.

1.14.2 The Overlapping Clustering

The *fcm* algorithm [11] is the prototype of the overlapping clustering algorithms. It minimizes an objective function of the form (1.16)

$$J_m(U, V) = \sum_{i=1}^{c} \sum_{k=1}^{n} (\mu_{ik})^m \|x_k - v_i\|^2 \tag{1.16}$$

where $U = [\mu_{ik}]$ is the fuzzy partition matrix, $\mu_{ik} \in [0, 1], \forall i, k$, represents the degree to which the ith pattern belongs to the kth class, and $1 \leq m \leq \infty$, being $m = 2$ the most popular choice for this parameter. Thus, the *fcm* clustering allows a

given pattern to belong to different classes but with different degree of membership. The user may simply decide to assign the ith pattern to the class represented by the $\max(\mu_{i,k}), k = 1, \ldots, c$ or to evaluate in more detail, when there is not a clear prevalence of some μ_{ik}. For this reason, the *fcm* algorithm is usually preferred to the *k-means* algorithm. The *fcm* algorithm is implemented in MATLAB® as the function *fcm*.

Unlike hierarchical clustering (see Sect. 1.14.3), the *k-means* and *fcm* algorithms creates a single level of clusters, which is usually the most appropriate choice in presence of large amounts of data. For this reason, for time series clustering, *k-means* and *fcm* clustering is often more suitable than hierarchical approaches.

1.14.3 The Hierarchical Clustering

The hierarchical clustering approaches, such as the Random Forest [12], whose basic principles are attributed to [13], works by grouping data objects into a tree of clusters. The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. There are generally two types of hierarchical clustering methods agglomerative and divisive. Agglomerative methods start by placing each object in its own cluster and then merge clusters into larger and larger clusters, until all objects are in a single cluster or until certain termination conditions such as the desired number of clusters are satisfied. Divisive methods do just the opposite. The function *clusterdata* supports agglomerative clustering and performs all the necessary steps.

1.14.4 The Probabilistic Clustering

The probabilistic clustering model-based approaches, such as the Mixture of Gaussians [14], assumes that the data set follows a mixture model of probability distributions so that a mixture likelihood approach to clustering may be used. For a mixture model, the expectation and maximization (EM) algorithm is commonly used, which assigns posterior probabilities to each component density with respect to each observation. Clusters are assigned by selecting the component that maximizes the posterior probability. Like *k-means* clustering, *Gaussian mixture* modeling uses an iterative algorithm that converges to a local optimum. *Gaussian mixture* modeling may be more appropriate than *k-means* clustering when clusters have different sizes and correlation within them. The posterior probabilities for each point indicate that each data point has some probability of belonging to each cluster. The *Gaussian mixture* clustering approach is implemented in the MATLAB® Statistics Toolbox, under the class *gmdistribution*.

1.14.5 Feature Based Clustering

As concerning the kinds of data considered, time series clustering methods can be classified into three major categories, depending upon whether they work directly with raw data, indirectly with features extracted from the raw data, or indirectly with models built from the raw data. Clustering patterns of time series, if the individual patterns consist of dozens or hundreds of values, the problem arises in a high-dimensional space and may be for this reason rather complex. For this reason, it is often not recommended [8]. Instead, in this case it is more appropriate the feature based approach, which means that original time series pattern is preliminary processed in order to extract a limited number of features, which are then used to perform the clustering. Examples of feature based clustering will be given in Chaps. 7 and 8 of this book.

1.14.6 Choosing the Number of Clusters

A common problem, dealing with unsupervised clustering, is that of choosing the most appropriate number of clusters. To this purpose various approaches have been proposed in literature, such as the Self-Organized Maps (SOM) [15], and the Silhouette [16] approaches.

1.14.6.1 Self-Organized Maps

SOM consists of neurons organized on a regular low-dimensional grid. The SOM can be thought as a network which is spread to space of input data. The network training algorithm moves the SOM weight vectors so that they span across the data such that neighboring neurons on the grid get similar weight vectors. After training the SOM it is possible to plot some useful feature, such as for instance the so-called hit matrix, with each neuron showing the number of input vectors that it classifies, as shown for instance in Fig. 1.3. In more detail, the Figure shows that essentially 3 of the 25 neurons of the 5 by 5 SOM network are those most concerned, thus meaning that the features considered as input, would be better classified into 3 clusters.

SOM networks can be configured by using the function *selforgmap*, trained by using the *train* function, while the hit matrix plot can be performed by using the *plotsomhits* function.

1.14.6.2 Silhouette

The silhouette $S(i)$ is a measure of how well the ith pattern lies within its cluster, and formally defined as (1.17).

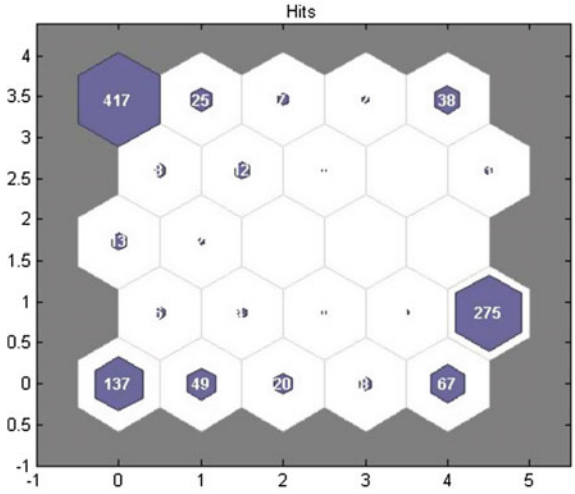

Fig. 1.3 Hit matrix obtained by training a 5 by 5 SOM network by using the Solar Radiation features at the station ID690140

$$S(i) = \frac{LD(i) - D(i)}{max\{D(i), LD(i)\}} \tag{1.17}$$

where $D(i)$ is the average dissimilarity of the ith pattern with all other data within the same cluster, and $LD(i)$ is the lowest average dissimilarity of the ith pattern to any other cluster, of which it is not a member. $D(i)$ is assumed as a measure of how well the ith pattern is assigned to its cluster (the smaller the value, the better the assignment). Thus an $S(i)$ close to 1 means that the corresponding pattern is appropriately clustered; on the contrary, If $S(i)$ is close to -1, then the ith pattern would be more appropriate if it was clustered in its neighboring cluster. An $S(i)$ near zero means that the pattern is on the border of two natural clusters.

Application of this approach to evaluate the consistency of clustering in 2 or more clusters solar radiation and wind speed daily patterns will be given in Chaps. 7 and 8, respectively.

Evaluate of clustering solutions by using the silhouette criterion can be performed by using the *evalclusters* function, while the silhouette plot can be performed by using the *silhouette* function.

1.15 Conclusions

In this chapter the main techniques that will be considered in this book to analyze the properties of solar radiation and wind speed time series were presented. The description has been limited to the essentials and then for further details it is possible to refer to textbooks such as [17].

References

1. R. Hegger, H. Kantz, T. Schreiber, Practical implementation of nonlinear time series methods: the TISEAN package. Chaos **9**, 413–435 (1999)
2. B. Mandelbrot, J. Wallis, Robustness of the rescaled range R/S in the measurement of noncyclic long-run statistical dependence. Water Resour. Res. **5**, 967–988 (1969)
3. R. Weron, Estimating long range dependence finite sample properties and confidence intervals. Physica A **312**, 285–299 (2002)
4. B.B. Mandelbrot, *The Fractal Geometry of Nature* (Macmillan, 1983)
5. J.W. Kantelhardt, S.A. Zschiegner, E. Koscielny-Bunde, A. Bunde, S. Havlin, H.E. Stanley, Multifractal detrended fluctuation analysis of nonstationary time series. Physica A **316**, 87–114 (2002)
6. C. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, A.L. Goldberger, Mosaic organization of DNA nucleotides. Phys. Rev. E **49**, 1–14 (1994)
7. E.A.F. Ihlen, Introduction to multifractal detrended fluctuation analysis in MATLAB. Front. Physiol. 1–18 (2012). http://dx.doi.org/10.3389/phys.2012.00141
8. T.W. Liao, Clustering of time series data–a survey. Pattern Recogn. **38**, 1857–1874 (2005)
9. S. Aghabozorgi, A.S. Shirkhorshidi, T.Y. Wah, Time-series clustering A decade review. Inf. Syst. **53**, 16–38 (2015)
10. J.B. MacQueen, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1 (University of California Press, 1967), pp. 281–297
11. J. Bezdek, R. Ehrlich, W. Full, FCM: the fuzzy c-means clustering algorithm. Comput. Geosci. **10**, 191–203 (1984)
12. L. Breiman, Random forests. Mach. Learn. **45**, 5–32 (2011)
13. S.C. Johnson, Hierarchical clustering schemes. Psychometrika **2**, 241–254 (1967)
14. G.J. McLachlan, K.E. Basford, *Mixture Models Inference and Applications to clustering* (Marcel Dekker, New York, 1988)
15. T. Kohonen, *Self-Organizing Maps* (1995)
16. P.J. Rousseeuw, Silhouettes a graphical aid to the interpretation and validation of cluster analysis. Comput. Appl. Math. **20**, 53–65 (1987). doi:10.1016/0377-0427(87)90125-7
17. J.C. Sprott, *Chaos and Time-Series Analysis* (2003)

Chapter 2
Analysis of Solar Radiation Time Series

Abstract In this chapter, various kinds of analysis are performed by using solar radiation time series recorded at 12 stations of the NREL database. Analysis are devoted to infer some basic properties, such as stationarity, autocorrelation, and the embedding phase-space dimension. These features will be considered in the rest of the book to implement short-term forecasting models and perform the clustering of solar radiation daily patterns. Furthermore, others kinds of analysis are carried out, such as the fractal and multifractal analysis and estimation of Lyapunov exponents, in order to clarify more deeply the nature of solar radiation time series.

2.1 Energy from the Sun

The Sun is certainly the main source of renewable energy. Just to have an idea it is possible to say that the Sun delivers toward the surface of the terrestrial hemisphere exposed a power exceeding 50 thousand Tera Watt which is about 10 thousand times the energy used all over the world [1]. A part of this energy reaches the outer part of the Earth's atmosphere with an average irradiance of about 1367 W/m^2, a value which varies as a function of the Earth-to-Sun distance and of the solar activity. The problem of estimating the global horizontal solar radiation $GHI(h, d)$, for any hour h of a day d of the year, at any site, has been addressed in literature by several authors such as [2]. It depends on a quite large number of parameters which, roughly speaking, can be summarized as follows: the distance from the sun, the duration of the daily sunlight period, the inclinations of solar rays to the horizon, the transparency of the atmosphere toward heat radiation and the output of solar radiation. Some of these factors are connected with mechanical parameters which describes the revolution of Earth around the Sun and on the Earth spinning about itself. Others factors depend on the properties of the atmosphere and are stochastic in nature, such as the cloud cover features (size, speed, and number) and the degree of pollution.

When passing through the atmosphere, the solar radiation decreases in intensity because it is partially reflected and absorbed, mainly by the water vapor and others atmospheric gases. The radiation which passes through is partially diffused by the air and by the solid particles suspended in the air. Therefore, the radiation falling

L. Fortuna et al., *Nonlinear Modeling of Solar Radiation and Wind Speed Time Series*, SpringerBriefs in Energy, DOI 10.1007/978-3-319-38764-2_2

on a horizontal surface is constituted by a direct radiation, associated to the direct irradiance on the surface, by a diffuse radiation and by a radiation reflected on a given surface by the ground and by the surrounding environment. In winter, when the sky is usually overcast, the diffuse component is greater than the direct one.

2.2 The Solar Radiation Data Set

The data set considered in this book consists of hourly average time series recorded at 12 stations stored in the USA National Solar Radiation Database managed by the NREL (National Renewable Energy Laboratory). Data of this database was recorded from 1999 to 2005 and can be freely download as described in Appendix A.2. The 12 stations were selected based on two criteria: the quality of time series and the need to ensure the necessary diversification of meteo-climatic conditions. The 12 selected stations are listed in Table 2.1. More detailed information about these and others recording stations of the National Solar Radiation Database can be found in [3]. Analysis of solar radiation time series has been addressed in literature by various authors such as [4–6]. Nevertheless, as the available results are sometimes fragmented, in this section it is provided a picture as comprehensive as possible of features of this kind of time series. Analysis performed refer to aspects such as stationarity, power spectrum, autocorrelation, fractal, and embedding state-space dimension. According with the basic knowledge about solar radiation, Fig. 2.1 shows that the considered kinds of time series are fluctuating at any time scale. Indeed, in the Figure a time series is shown at hourly, daily, monthly, and yearly time scales. Fluctuations observed in solar radiation time series is a feature shared with other meteorological time series, such as wind speed. These fluctuations are superimposed

Table 2.1 Coordinates of the 12 solar radiation recording stations

stationID	Description	*Lat*	*Lon*	*Elev*	*UTC*
690140	El Toro MCAS (CA)	33.667	−117.733	116	−8
690150	Twenty-nine Palms (CA)	34.3	116.167	626	−8
722020	Miami Int AP (FL)	25.817	−80.3	11	−5
722350	Jackson Int AP (MS)	32.317	−90.083	94	−6
722636	Dalhart Municipal AP (TX)	36.017	−102.55	1216	−6
723647	Albuquerque Double (NM)	35.133	−106.783	1779	−7
724776	Moab Canionlands(UT)	38.58	−109.54	1000	−7
725033	New York Central PRK (NY)	40.783	−73.967	40	−5
725090	Boston Int AP, (MA)	42.367	−71.017	6	−5
726055	Pease Int Tradepor (NH)	43.083	−70.817	31	−5
726130	Mount Washington (NH)	44.267	−71.3	1910	−5
726590	Aberdeen Regional AP (SD)	45.45	−98.417	398	−6

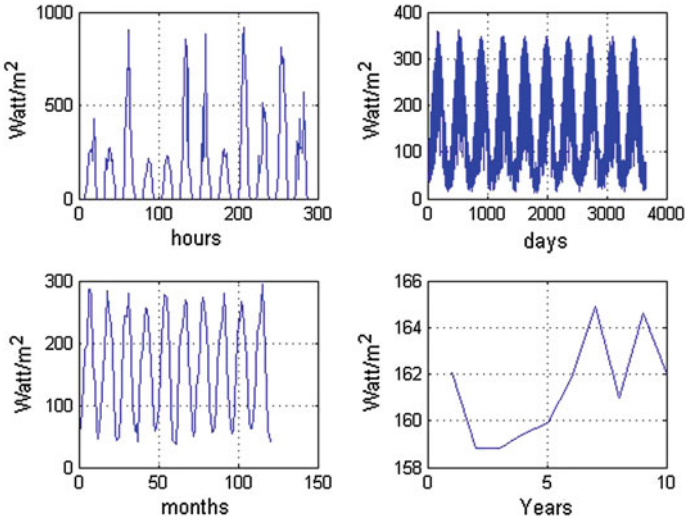

Fig. 2.1 Solar radiation time series at hourly, daily, monthly, yearly scale, respectively

with deterministic variations due to the Earth spinning around itself and to the revolution of Earth around the Sun. The Earth spinning determines the typical bell shape curves that are visible at hourly scale, while the revolution around the Sun determines the fluctuations that are visible at monthly scales. However, fluctuations occur also from year to year, as shown in lower rightmost sub Fig. 2.1.

2.2.1 Stationarity Analysis

One of the preliminary issues that one would like to know is if solar radiation time series are stationary. To this purpose, usually, available tests are based on the search for existent of a *unit root*, such as the Dicky–Fuller and the Phillips–Perron tests (see the *adftest* and the *pptest* functions, listed in Appendix A.1). The application of these tests indicate that the null-hypothesis, i.e., the considered time series are nonstationary, is false.

Another approach to search for nonstationary evidences in time series is that of using recurrent plots (see the function *recurr* in Appendix A.1). Such a kind of plots obtained from hourly average solar radiation time series, for two different embedding dimensions, are shown in Fig. 2.2. Since it is known that in an ergodic situation, the dots of a recurrent plot should cover, on average, the plane uniformly, whereas nonstationarity expresses itself by an overall tendency of dots to be close to the diagonal, it is possible to conclude that there are not evidences that solar radiation time series are nonstationary, at least for time intervals of 10 years, as for the data set considered in this book.

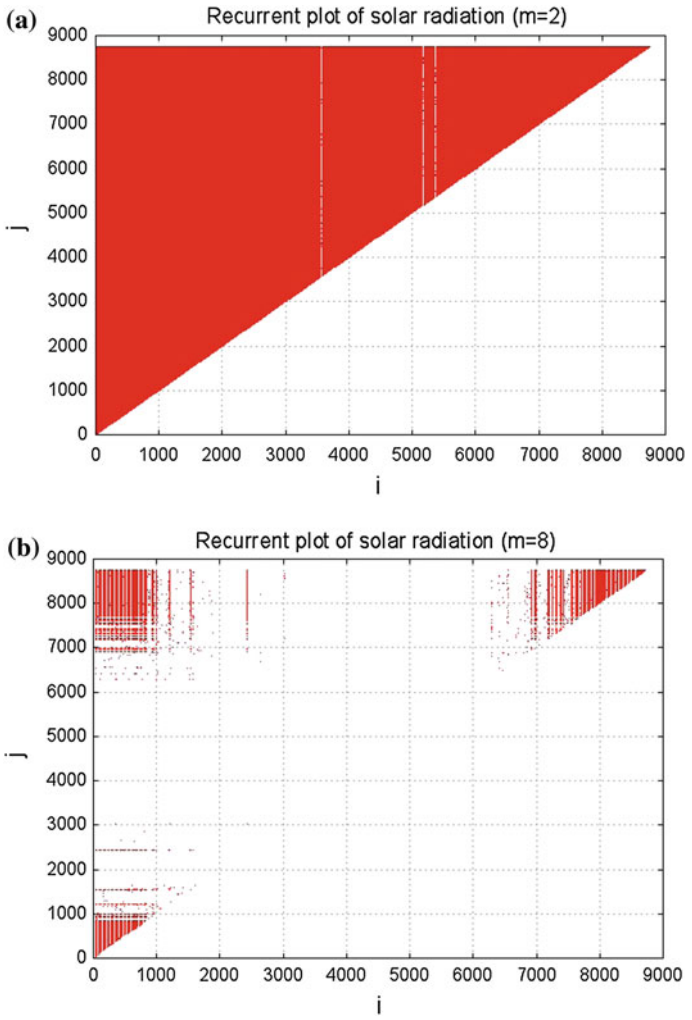

Fig. 2.2 Recurrent plots of solar radiation for two different embedding dimensions. **a** m = 2. **b** m = 8

2.2.2 *Autocorrelation and Mutual Information*

The autocorrelation functions computed for hourly and daily average solar radiation time series are reported in Fig. 2.3. As it is possible to see the autocorrelation function of hourly average solar radiation is strongly periodic with period 24 h, due to the marked daily component, already pointed out (see the upper leftmost Fig. 2.1). Furthermore, the autocorrelation function computed at hourly scale decays at values lower than 0.37, the so-called correlation time τ_c, in about 5 h. At daily scale the

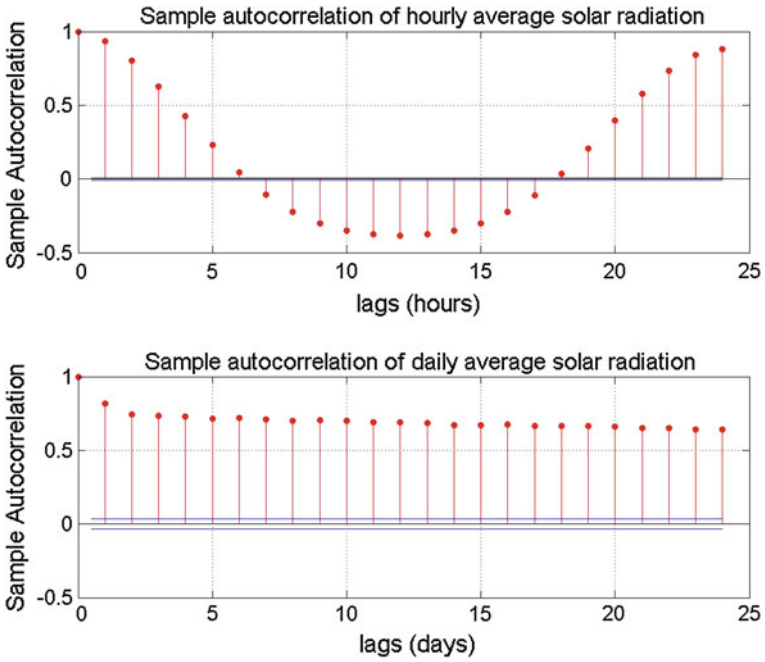

Fig. 2.3 Autocorrelation of typical hourly and daily average solar radiation time series

autocorrelation decays very slowly, which means that daily average solar radiation time series are long-range correlated.

Since the autocorrelation is a linear feature of time series, while it is reasonable that nonlinear processes are involved with solar radiation, it is useful to estimate also the mutual information, as shown in Fig. 2.4. A popular rule, referred to as the first minimum criterion, usually considered in evaluating the mutual information, is that two samples can be considered statistical independent if they are delayed by a number of samples equal to the time needed for the mutual information to reach the first minimum. By using this criterion at hourly scale two samples can be considered statistically independent if delayed by about six lags, while at daily scale if delayed about one lag. Thus, the analysis of the mutual information essentially confirms the results gathered by the autocorrelation function at hourly scale, but provides some more insight at daily scale.

2.2.3 Power Spectra

The typical power spectra of hourly and daily solar radiation time series are shown in Fig. 2.5. It is possible to observe that at hourly scale there are marked components with periods:

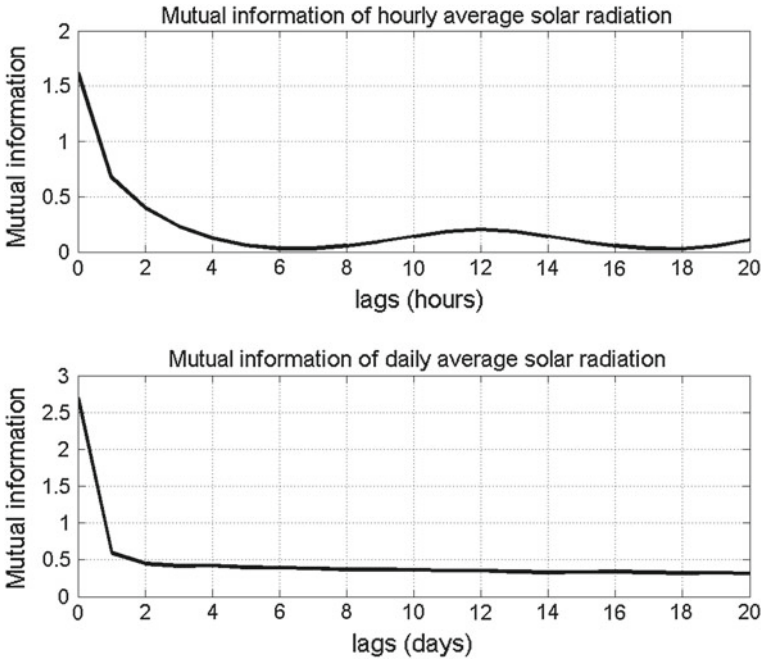

Fig. 2.4 Mutual information of typical hourly and daily average solar radiation time series

$T_1 = 1/0.0001143 \simeq 8748\,\text{h} \simeq 1\,\text{year}$,
$T_2 = 1/0.04167 \simeq 24\,\text{h}$.

The others components of the spectrum computed at hourly scale corresponding to periods of 12 h, 6 h etc., are well-known effects of the considered fast Fourier transform (FFT) computing algorithm. At daily scale only one marked component is evident, corresponding to a period of $T_3 = 1/0.002743 \simeq 365$ days, i.e., 1 year.

The absolute values of the power spectra slopes computed for all the stations considered in this book are reported in Table 2.2. We obtained that the mean slope of hourly and daily time series is about 1.05 and 0.67 respectively. The difference among these slopes can be easily explained bearing in mind that daily average solar radiation time series are less autocorrelated than the corresponding hourly average, and thus, more similar to a white noise. However, both hourly and daily average time series can be classified as $1/f$ noises (see Sect. 1.8).

2.2.4 Hurst Exponent and Fractal Dimension

The Hurst exponent of hourly average time series obtained by using the detrended fluctuation algorithm (DFA) [7], and the fractal dimension, computed by using the

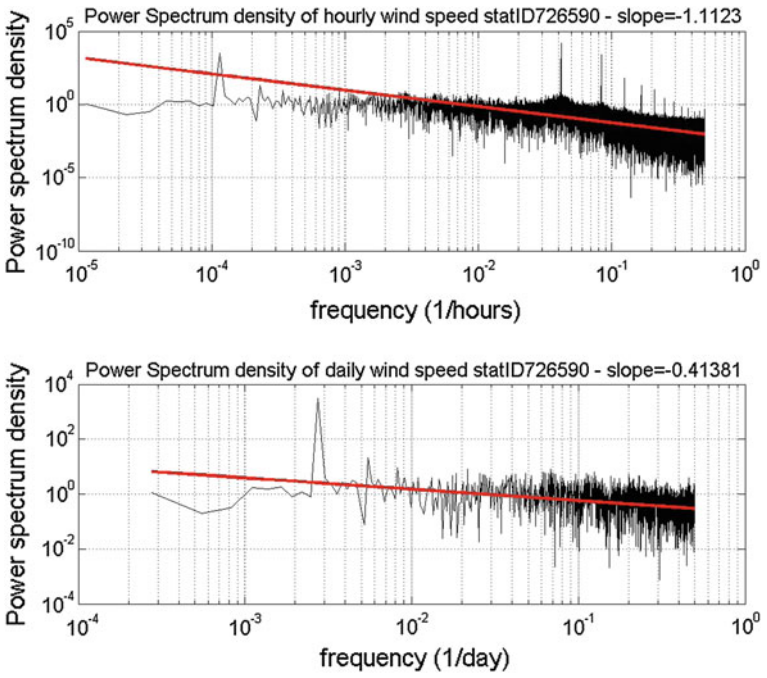

Fig. 2.5 Power spectrum densities of hourly and daily average solar radiation time series at the station ID 726590

Table 2.2 Absolute slopes of power spectra of hourly (column 2) and daily (column 3) solar radiation averages, Hurst exponent (column 4), and box dimension (column 5) of the hourly solar radiation time series at the 12 considered stations

stationID	$\beta(hourly)$	$\beta(daily)$	H	D
690140	1.1651	0.8443	0.7614	1.4603
690150	0.8992	0.8928	0.7806	1.4601
722020	0.8236	0.7104	0.6740	1.4642
722350	1.0973	0.6962	0.7326	1.4599
722636	1.1147	0.6664	0.7655	1.4592
723647	0.9262	0.6506	0.7549	1.4601
724776	0.8753	0.6454	0.7965	1.4638
725033	1.1706	0.5538	0.7661	1.4618
725090	1.2606	0.5329	0.7740	1.4603
726055	1.1999	0.5682	0.7742	1.4597
726130	1.0360	0.6198	0.7951	1.4630
726590	1.0932	0.6837	0.8258	1.4604

boxcounting algorithm [8] at the considered stations are shown in Table 2.2. It is possible to see that, H and D, gives on average $H = 0.77$ and $D = 1.46$. These values clearly point out the fractal nature of solar radiation time series. In particular, the Hurst exponent indicates that they are positive long-range correlated since $0.5 < H < 1$ (see Sect. 1.9.2).

2.2.5 Multifractal Analysis of Solar Radiation

One of the techniques proposed in the literature to deal with multifractal is the multifractal detrended fluctuation analysis (MFDFA) [9] developed as the extension of the DFA. In the framework of MFDFA analysis a generalized Hurst exponent $H(q)$ is defined, depending on the local fluctuation exponent q. Furthermore, it is defined the so-called multifractal spectrum (also referred to as singularity spectrum). According with [6], who recognized the multifractal nature of solar radiation time series, we report some further evidences obtained by performing the multifractal analysis at some of the stations listed in Table 2.1. The generalized Hurst exponents and the corresponding multifractal spectrum of these stations are reported in Fig. 2.6a and b, respectively. Figure 2.6a clearly shows a marked dependence of $H(q)$ on q, thus confirming the multifractal nature of solar radiation time series. In order to interpret the multifractal spectrum reported in Fig. 2.6b it is to bearing in mind that it has an asymmetric bell shape with the maximum obtained for $q = 0$. The multifractal spectrum will have a long left tail when the time series have a multifractal structure that is insensitive to the local fluctuations with small magnitude. In contrast, the multifractal spectrum will have a long right tail when the time series have a multifractal structure that is insensitive to the local fluctuations with large magnitudes. Figure 2.6b shows that the five stations exhibit right tails, i.e., the spectra are left truncated, which simply means that while large fluctuation scales within a limited range of exponents $\alpha \in [0.66, 0.7]$, the small fluctuation scales following a wider range of exponents $\alpha \in [0.7, 1]$.

2.2.6 Estimation of the Embedding Dimension

In order to determine the embedding dimension d of the dynamical system underlying solar radiation time series, the fraction of false nearest neighbors was computed, as shown for instance in Fig. 2.7. The Figure shows that the fraction of false nearest neighbors decays very slowly with the embedding dimension, without reaching the zero value in the range $d \in [1, 30]$. However, it seems that the curve reaches a steady state for $d \geq 24$. While the fact that a zero value is not reached can be interpreted as the effect of noise in the considered time series, the value $d = 24$ can be justified bearing in mind that hourly average solar radiation time series exhibits a strong daily component, i.e., with period 24 h.

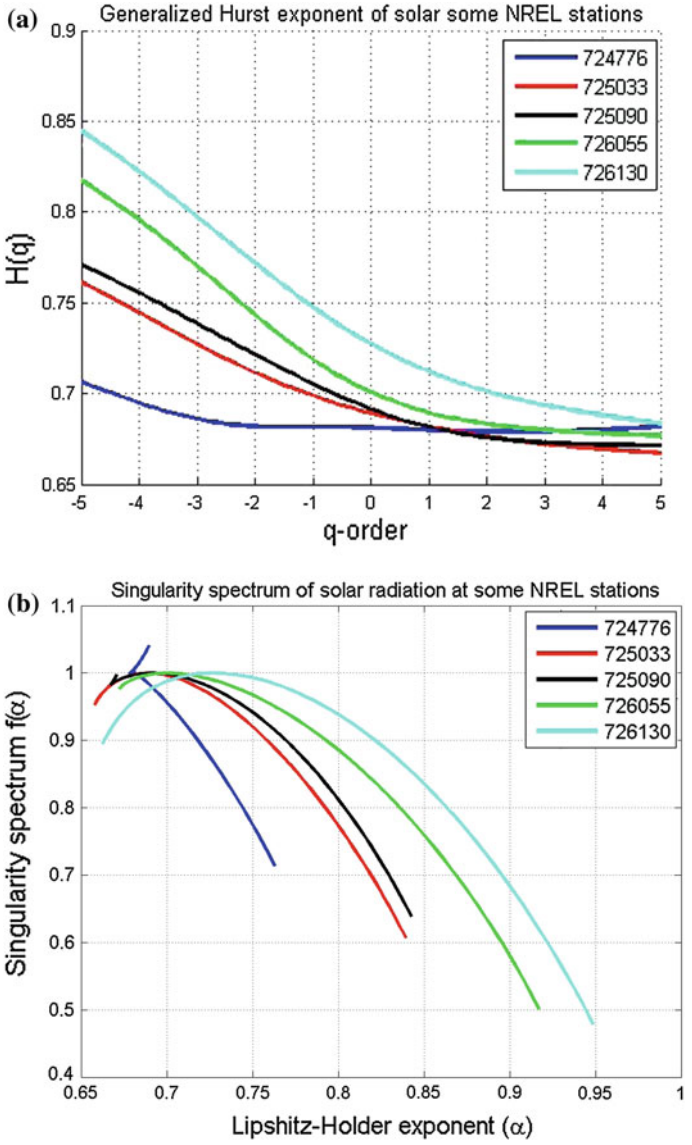

Fig. 2.6 Generalized Hurst exponent and singularity spectrum at some of the considered solar radiation recording stations. **a** Generalized Hurst exponent. **b** Singularity spectrum

2.2.7 Maximal Lyapunov Exponent

As mentioned in Sect. 1.12, a bounded dynamical system with a positive Lyapunov exponent is chaotic, and the so-called maximal exponent describes the average rate at which predictability is lost (see [10]). Results obtained computing the maximal Lya-

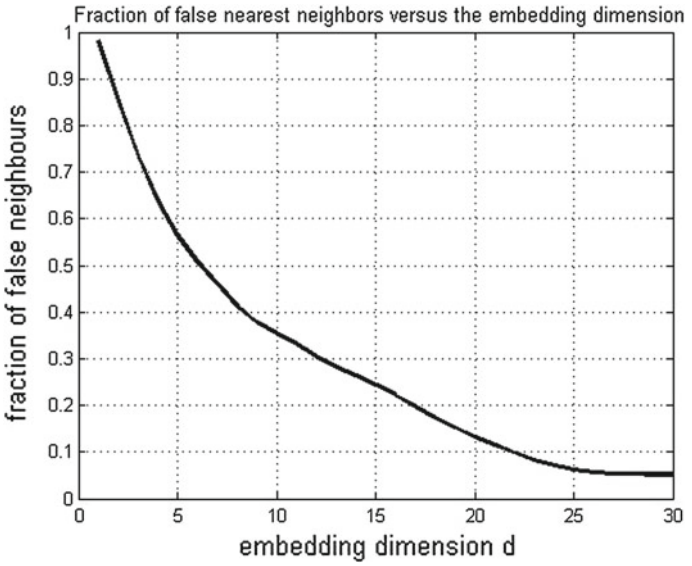

Fig. 2.7 Fraction of false nearest neighbors of solar radiation versus the embedding dimension d

punov exponent of hourly average time series recorded at the considered stations, by using the *lyap_k* function indicate the existence of positive maximal exponents in the range [0.6, 1]. This result would indicate that the hourly average solar radiation time series are chaotic. However, since this result if confirmed, would have some interesting implications, it is given here as a work in progress and further investigations are needed.

2.3 Conclusions

Analysis presented in this Chapter, performed on both hourly and daily average solar radiation time series allow to draw some conclusions about their nature. Stationary analysis, carried out by different approaches, has not pointed out evidences that they are nonstationary, at least for time intervals of 10 years, which is the largest considered in this study. The power spectrum analysis has shown that the slopes of the solar radiation time series power spectra are in the range [0.5, 1.5], which means that solar radiation time series belong to the wide class of $1/f$ noise. Correlation analysis, carried out by using linear and non linear approaches, pointed out that solar radiation time series exhibits a correlation time of about $\tau_c = 5$ lags at hourly time scale and of about $\tau_c = 1$ lag at daily scale, which means that prediction models, based on autocorrelation, can be reliable for short horizons only. Fractal analysis pointed out that these kind of time series are fractal exhibiting, on average, fractal

dimension $D = 1.46$ and Hurst exponent of $H = 0.77$. Furthermore, the multifractal detrended fluctuation analysis (MFDFA) has pointed out that solar radiation time series are multifractal. analysis carried out in order to see if there are evidences of deterministic chaos, as suggested by the presence of at least a positive exponent in the range [0.6, 1], needs further investigations, since as well known the computation of Lyapunov exponents from time series is rather difficult and may have some drawbacks.

References

1. AAVV, Photovoltaic plants, ABB Technical Application Paper N. 10, 1–124
2. S. Kaplanis, E. Kaplani, A model to predict expected mean and stochastic hourly global solar radiation values. Renew. Energy **32**, 1414–1425 (2007)
3. S. Wilcox, National Solar Radiation Database 1991-2010 Update—Users Manual, Technical Report NREL/TP-5500-54824, 1–479, 2012. ftp://ftp.ncdc.noaa.gov/pub/data/nsrdb-solar/documentation-2010/
4. A. Maafi, S. Harrouni, Preliminary results of the fractal classification of daily solar irradiance. Solar Energy **75**, 53–61 (2003)
5. S. Harrouni, A. Guessoum, Using fractal dimension to quantify long-range persistence in global solar radiation. Chaos Solitons Fractals **41**, 1520–1530 (2009)
6. Z. Zeng, H. Yang, R. Zhao, J. Meng, Nonlinear characteristics of observed solar radiation data. Solar Energy **87**, 204–218 (2013)
7. R. Weron, Estimating long range dependence finite sample properties and confidence intervals. Physica A **312**, 285–299 (2002)
8. N. Sarkar, B.B. Chaudhuri, An efficient differential box-counting approach to compute fractal dimension of image. IEEE Trans. Syst. Man Cybern. **24**, 115–120 (1994)
9. J.W. Kantelhardt, S.A. Zschiegner, E. Koscielny-Bunde, A. Bunde, S. Havlin, H.E. Stanley, Multifractal detrended fluctuation analysis of nonstationary time series. Physica A **87–114**, 316 (2002)
10. J.C. Sprott, *Chaos and Time-Series Analysis* (2003)

Chapter 3
Analysis of Wind Speed Time Series

Abstract In this chapter various kinds of analysis are performed by using wind speed time series recorded at 12 stations of the WWR database. Analyses are devoted to infer some basic properties, such as stationarity, autocorrelation, and the embedding phase-space dimension. These features will be considered in the rest of the book to implement short-term forecasting models and perform the clustering of wind speed daily patterns. Furthermore, others kinds of analysis are carried out, such as the fractal and multifractal analysis and estimation of Lyapunov exponents, in order to clarify more deeply the nature of wind speed time series.

3.1 Energy from the Wind

The wind is a quite complex phenomenon primarily generated by heating and cooling processes naturally and continuously occurring in the lower atmosphere. The main factors that affect wind speed and direction are the pressure-gradient force, the Coriolis force, and friction with the Earth surface. The pressure-gradient force is ultimately generated by the unevenly way in which the Earth continuously releases into the atmosphere the heat received by the Sun. In the areas where less heat is released the pressure of atmospheric gases increases, whereas where more heat is released, air warms up and gas pressure decreases. As a consequence, a macrocirculation due to the convective motions is created. In other terms, the different ways of receiving and releasing the heat received from the Sun generate air movements from areas where the atmospheric pressure is higher toward areas where it is lower. Therefore, wind is the movement of an air mass, more or less quick, between zones at different pressures. The greater the pressure difference, the quicker the airflow and consequently the stronger the wind. The Coriolis force, due to the Earth spinning, is a force that essentially influence the direction of the winds. It explains why wind does not blow in the direction joining the center of the high pressure with that of the low pressure, but in the northern hemisphere it veers to the right, circulating around the high pressure centers with clockwise rotation and around the low pressure ones

L. Fortuna et al., *Nonlinear Modeling of Solar Radiation and Wind Speed Time Series*, SpringerBriefs in Energy, DOI 10.1007/978-3-319-38764-2_3

in the opposite direction. In the practice, who keeps his back to the wind has on his left the low pressure area and on his right the high pressure area. In the southern hemisphere the opposite occurs.

Friction is the third force that affects both speed and direction of winds. It essentially operates in the proximity of the Earth's surface, but, nevertheless it influences the force balance, reducing the Coriolis force in proximity of the ground and moving the air at right angles across the isobars toward the area of lower pressure. On a large scale, at different latitudes, a circulation of air masses can be noticed, which is cyclically influenced by the seasons. On a smaller scale, there is a different heating between the dry land and the water masses, with the consequent formation of the daily sea and Earth breezes. The profile and unevenness of the surface of the dry land or of the sea deeply affect the wind and its local characteristics; in fact, the wind blows with higher intensity on large and flat surfaces, such as the sea: this represents the main element of interest for wind plants on and off shore. Moreover, the wind gets stronger on the top of the rises or in the valleys oriented parallel to the direction of the dominant wind, whereas it slows down on uneven surfaces, such as towns or forests, and its speed with respect to the height above ground is influenced by the conditions of atmospheric stability [1].

From this description, despite short and rough, it should be clear that wind is a complex natural process, as pointed out also by analysis, carried out in the rest of this chapter.

3.2 The Wind Speed Data Set

The data set of wind speed time series analyzed in this chapter is a subset of the Western Wind Resource (WWR) database which stores data of more than 30, 000 sites that were modeled in the framework of Western Wind and Solar Integration Study [2]. Data are at the present public available (see Appendix A.2 for details) through a friendly user interface. Original time series consists 10 min averages of the wind speed measured, from 2004 to 2006 on aeolic towers at 100 meters from the ground. The database includes also data about electric power generated by the aeolic plants and geographic coordinates (Latitude, Longitude, and altitude) of recording sites. The subset of stations considered in this study are listed in Table 3.1. The selected stations were chosen in different geographic areas and altitude with the aim of preserving the generality of results. As an example, the time series recorded at the station ID2257 are shown in Fig. 3.1, which demonstrate that wind speed, such as solar radiation, are fluctuating time series at any timescale.

Table 3.1 Coordinates of the 12 wind speed recording stations

StationID	State	*Lat* (N)	*Lon* (W)	*Elev*(*m*)
2257	CA	34.58	120.67	0
2300	CA	34.59	120.66	19
6435	AZ	36.14	113.07	1782
9004	CA	37.64	118.84	2201
9210	CA	37.73	118.96	2357
9390	CA	37.81	119.01	2449
11240	NV	39.48	114.14	1871
11651	NV	39.67	114.09	2156
12684	NV	40.28	114.19	1775
13562	NV	40.98	114.56	1993
18993	ID	42.01	113.82	1873
25766	WY	44.88	109.14	1324

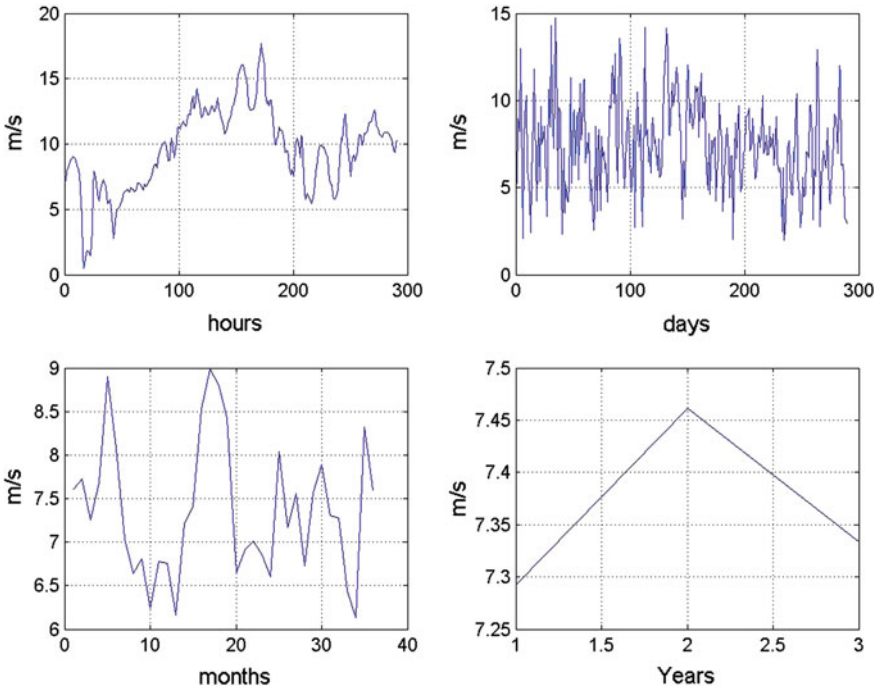

Fig. 3.1 Wind speed time series at hourly, daily, monthly, yearly scale, respectively

3.3 Stationary Analysis

Stationary analysis was performed by various techniques, including the Augmented Dicky-Fuller (ADF) test, the Phillips Pearson (PP) test, and the variance ratio (VR) test. All test rejected the null hypothesis, thus meaning that there are not enough evidences to assess that the considered time series are nonstationary, at least for timescale of 3 years.

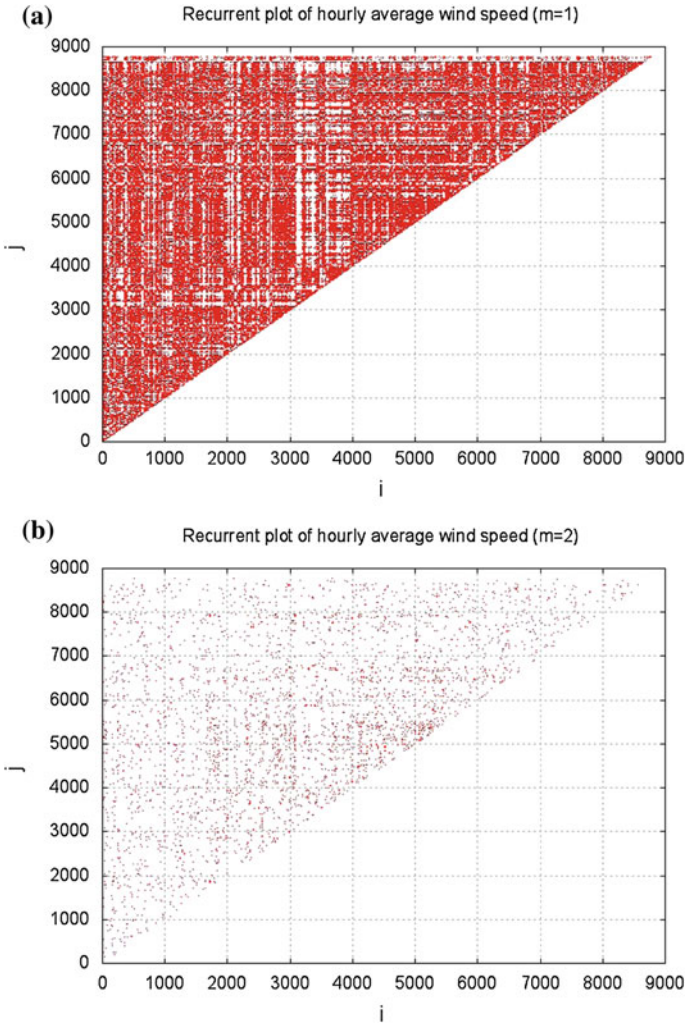

Fig. 3.2 Recurrent plots of hourly wind speed at the station ID 2257 during 2004 for different embedding dimensions **a** m = 1, **b** m = 2

Stationarity was also tested by considering the recurrent plots of hourly wind speed time series, computed for two different embedding dimensions (see Fig. 3.2). The uniform distribution of dots in the recurrent plots indicates that there are not particular structures that could be related with non-stationarity, thus confirming achievement of the ADF, PP, and VR tests.

3.4 Autocorrelation and Mutual Information

The power spectrum and autocorrelation function computed on hourly average wind speed time series are reported in Fig. 3.3. As it is possible to see at hourly scale the autocorrelation function exhibits a slow decaying behavior, which is typical of $1/f$ noise. Indeed, autocorrelation at daily scale decays at meaningless levels after 2 lags. Autocorrelation of wind speed time series was also estimated in terms of mutual information, as shown in Fig. 3.4. The figure, in essence, shows that the correlation time τ_c is about $5 \div 6$ lags at hourly scale and $1 \div 2$ lags at daily scale.

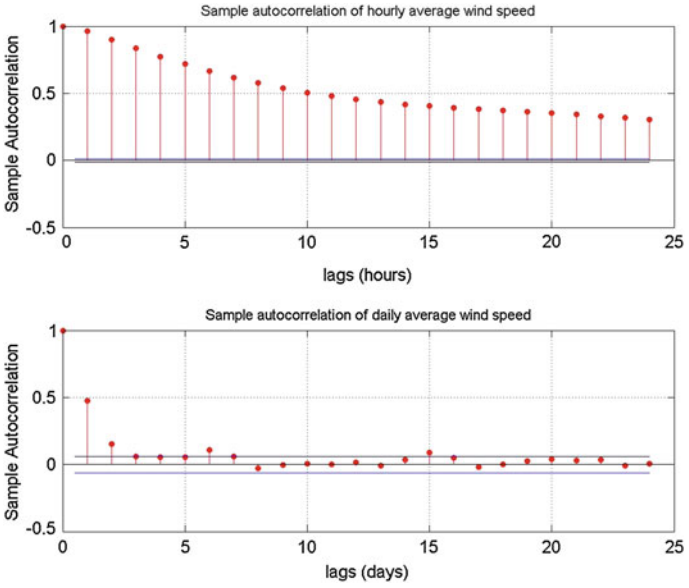

Fig. 3.3 Autocorrelation of hourly and daily average solar radiation time series at the ID2257 station

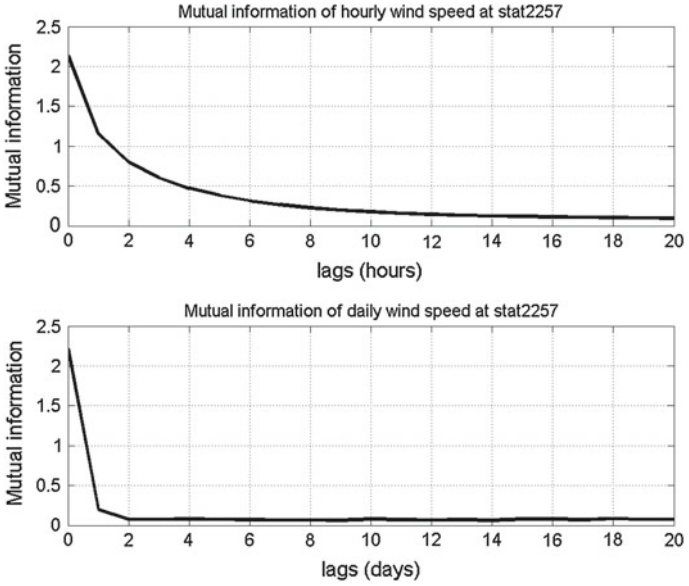

Fig. 3.4 Mutual information of hourly and daily average wind speed time series at the ID2257 station

3.5 Power Spectra

Typical power spectra of hourly and daily average wind speed time series are shown in Fig. 3.5. It is possible to observe that at hourly scale there are components with periods: $T_1 = 1/0.0001143 \simeq 8748\,\text{h} \simeq 1$ year, $T_2 = 1/0.04167 \simeq 24$ h. At daily scale only one component is evident corresponding to a period of $T_3 = 1/0.002743 \simeq 365$ days, i.e., 1 year.

In order to characterize the daily component, we have averaged hourly wind speed as described in Sect. 1.13. As an example, wind speed daily patterns computed for the station ID 2257 are shown in Fig. 1.1.

Wind speed patterns can be observed also at yearly timescale, as shown in Fig. 3.6. The shape of these patterns depends on the particular station, but is typically clearly recognizable.

The absolute slopes of hourly and daily average wind speed power spectra computed the considered stations are reported in Table 3.2. As it is possible to see $\beta \in [1.65, 2]$ and in $\beta \in [0.6, 1]$ at hourly and daily scale, respectively. This means that wind speed time series perform as $1/f$ noise or as random walks. The difference in slope between hourly average wind speed and the daily average wind speed can be explained bearing in mind that the latter are less autocorrelated and therefore more similar to a white noise.

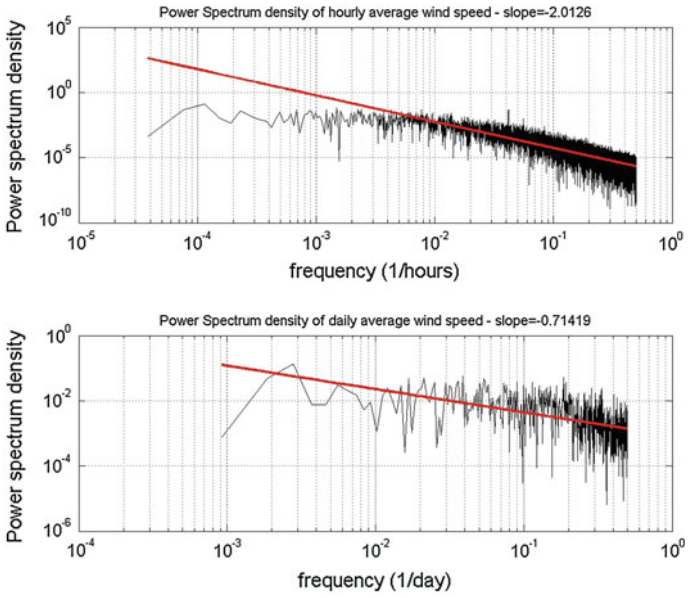

Fig. 3.5 Power spectrum densities of hourly and daily average solar radiation time series at the station ID2257

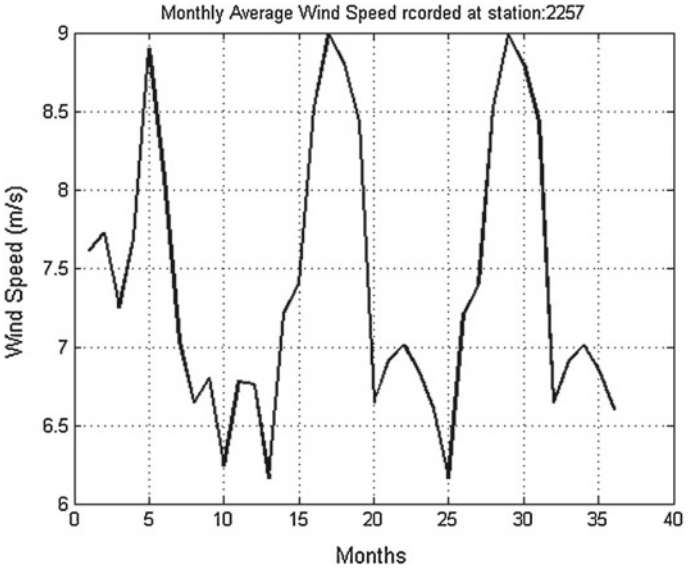

Fig. 3.6 Monthly average wind speed pattern from 2004 to 2006 at the station ID2257

Table 3.2 Absolute slopes of hourly and daily average power spectra at the considered stations

Station ID	$\beta\,(hourly)$	$\beta\,(Daily)$
2257	2.0258	0.96459
2300	2.0397	0.9298
6435	1.8683	0.74376
9004	1.8399	0.78385
9210	1.8627	0.9145
9390	1.8852	0.96777
11240	1.8311	0.69659
11651	1.778	0.76796
12684	1.8223	0.83475
13562	1.8702	0.79605
18993	1.8944	0.89359
25766	1.6457	0.61789

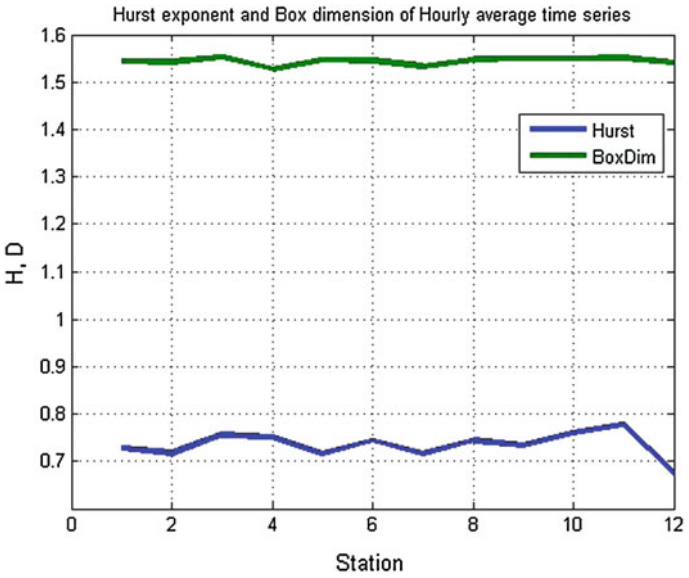

Fig. 3.7 Hurst exponent and fractal (box counting) dimension at the 12 considered stations

3.6 Hurst Exponent and Fractal Dimension

The Hurst exponents and the fractal dimensions of hourly average wind speed computed for each of the 12 considered stations are shown in Fig. 3.7. The Hurst exponent shown in this figure was computed by using the R/S algorithm while the fractal dimension was computed by using the box-counting algorithm. It is possible to see

that on average the Hurst exponent is 0.73 while the fractal dimension is 1.54. Thus the Hurst exponent is in the range 0.73 ± 0.09, observed for several natural systems [3].

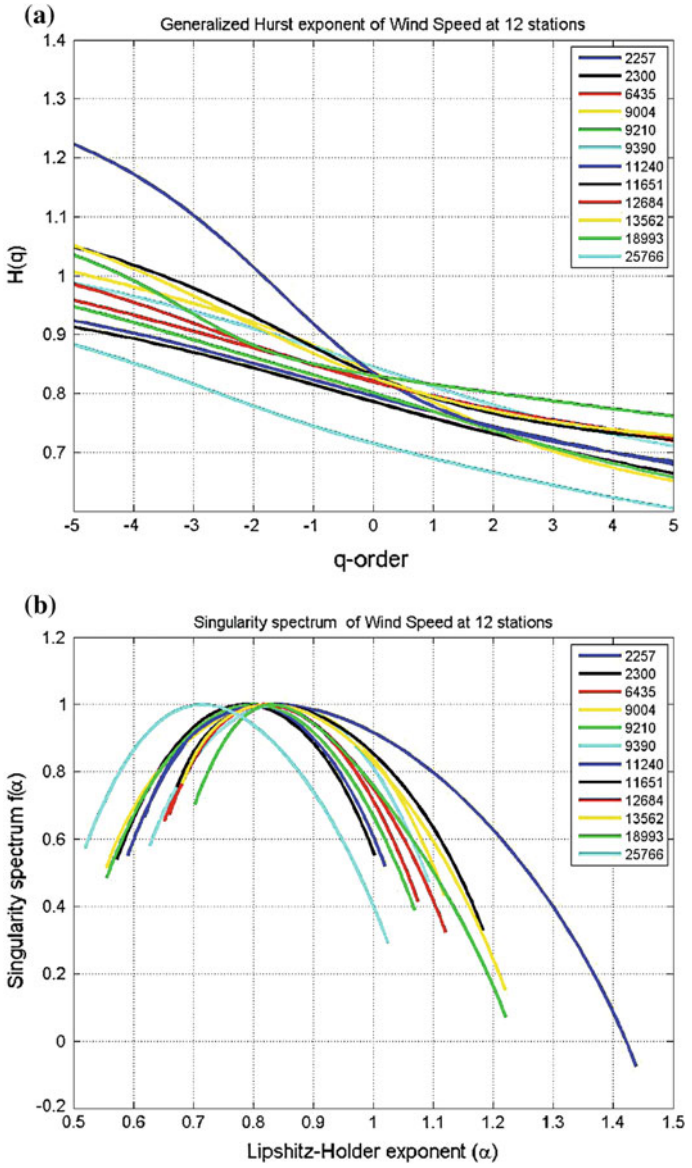

Fig. 3.8 Generalized Hurst exponent and singularity spectrum at 12 stations of the NREL database **a** generalized Hurst exponent **b** singularity spectrum

3.7 Multifractal Spectrum

The generalized Hurst exponent and the corresponding multifractal spectrum at 12 wind speed recording stations of the WWR database are shown in Fig. 3.8. Figure 3.8a shows that the generalized Hurst exponent $H(q)$ significantly varies versus q, clearly showing the multifractal nature of wind speed time series at all the considered stations, independently on the geographical area and altitude. In particular, it is possible to see that $H(2)$, i.e., the Hurst exponent computed by the traditional monofractal detrended fluctuation analysis (DFA), is in the range [0.65, 0.75]. Figure 3.8b shows that the singularity spectra are significantly affected by the different operating conditions of wind speed recording stations.

3.8 Estimation of the Embedding Dimension

In order to determine the embedding dimension of wind speed time series, the fraction of false nearest neighbors versus the embedding dimension was estimated, as shown, for instance, in Fig. 3.9. The figure shows that, similarly to what observed for solar radiation time series, the fraction of false nearest neighbors decays slowly with the

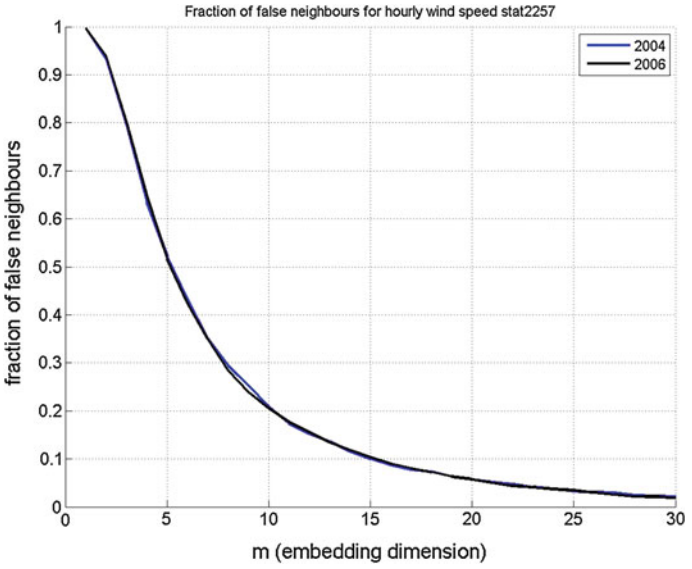

Fig. 3.9 Fraction of false nearest neighbors of wind speed at the station ID2257 during different years

embedding dimension and a small fraction is computed also for $d = 24$. This results may be due both to the presence of the daily component pointed out in Sect. 3.5 and to noise effecting the data set.

3.9 Lyapunov Exponents

The Lyapunov exponents, computed considering hourly average wind speed time series recorded at 12 stations of the WWR dataset, by using the *lyap_spec* function, assuming an embedding dimension $d = 24$, are shown in Fig. 3.10.

First of all, it is possible to see that the Lyapunov spectrum, is virtually identical for all stations, i.e., it is almost independent from the recording site. Furthermore, it is possible to observe that not only the greatest Lyapunov exponent is positive, but also about 10 others exponents. The value of about 0.12 estimated for the largest Lyapunov exponent was also obtained by using the *lyap_k* function, specifically designed for this purpose in the framework of the TISEAN project [5]. Therefore, this analysis seems to agree with [4] who pointed out the chaotic nature of wind speed time series.

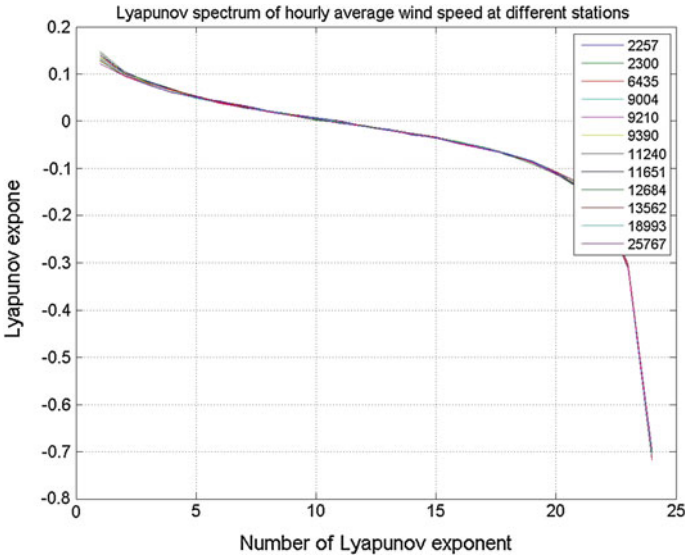

Fig. 3.10 Lyapunov spectrum of hourly average wind speed at 12 different recording stations of the WWR dataset

3.10 Conclusions

Analysis presented in this chapter, performed on both hourly and daily average wind speed time series allow to draw some conclusions about their nature. Stationary analysis, carried out by using different approaches, has not pointed out evidences that they are nonstationary, at least in time interval of 3 years, as analyzed in this book. Fractal analysis pointed out that this kind of time series is fractal and, in more detail, multifractal. The calculation of Lyapunov exponents, carried out at different stations, using two different functions, has highlighted the existence of chaos in these kinds of time series. However, this aspect and its implications require further study. Finally, the spectral analysis univocally indicates that wind speed time series belongs to the class of $1/f$ noise or random walks.

References

1. AAVV, Wind power plants, ABB Technical Application Paper N. 131–136
2. Western Wind and Solar Integration Study: The National Renewable Energy Laborator, (2010). http://www.nrel.gov/docs/fy10osti/47434.pdf
3. J.C. Sprott, Chaos and Time-Series Analysis (2003)
4. T.E. Karakasidis, A. Charakopoulos, Detection of low-dimensional chaos in wind time series. Chaos, Solitons and Fractals **41**, 1723–1732 (2009). doi:10.1016/j.chaos.2008.07.020
5. R. Hegger, H. Kantz, T. Schreiber, Practical implementation of nonlinear time series methods: the TISEAN package. Chaos **9**, 413 (1999)

Chapter 4
Prediction Models for Solar Radiation and Wind Speed Time Series

Abstract This chapter describes the structure of NARX and EPS models considered in this book to perform short-term prediction of solar radiation and wind speed time series. Furthermore, tools to identify the model parameters using both the adaptive neuro-fuzzy inference system (ANFIS) and the feed-forward neural networks (FFNN) approaches are outlined. Finally, it is described how model performances can be objectively assessed.

4.1 NARX Time Series Models

A time series is a sequence of measurements $y(t)$ of an observable y, performed at equal time intervals, which can be assumed as the output of an unknown dynamical system. The problem of modeling $y(t)$ consists in associating to the time series an appropriate dynamical model. In other words, it is the inverse of analyzing a dynamic system, after known its representation in terms of a set of differential equations, usually expressed in the so-called state space form.

The Takens theorem [1] implies that for a wide class of deterministic systems, there exists a diffeomorphism, i.e., a one-to-one differential mapping, between a finite window of the time series $((y(t), y(t-1), \ldots, y(t-d+1))$ and the state of the dynamic system underlying the series.

This implies that in theory there exist a MISO (multi-input-single-output) mapping $f : R^d \rightarrow R$ such that

$$y(t+1) = f(y(t), y(t-1), \ldots, y(t-d+1)) \tag{4.1}$$

where d (dimension) is the number of considered past values. This formulation returns a state space description where, in the d dimensional state space, the time series evolution is a trajectory and each point represents a temporal pattern of length d. Such a kind of prediction models are referred to as NAR (nonlinear autoregressive). Expression (4.1) generalizes into the so-called NARX (nonlinear autoregressive with exogenous, i.e., external, inputs) models [2, 3], represented by expression (4.2)

L. Fortuna et al., *Nonlinear Modeling of Solar Radiation and Wind Speed Time Series*, SpringerBriefs in Energy, DOI 10.1007/978-3-319-38764-2_4

$$y(t+1) = f(y(t), \ldots, y(t-d+1), u(t), \ldots, u(t-q+1)) \tag{4.2}$$

in presence of a vector $u(t)$ of explaining variables (or exogenous inputs), i.e., variables that are in some way correlated with $y(t)$ and fictitiously assumed as inputs of the dynamical model.

NAR and NARX have a linear counterpart into AR and ARX models.

4.2 Multistep Ahead Prediction Models

The problem of multistep ahead prediction is that of estimating $y(t+h)$, using information measured until time t. Here h is a positive integer, referred to as the prediction horizon. This problem can be solved in two ways: one-step prediction and iterated prediction. In the former case, expression (4.2) is rewritten as (4.3)

$$y(t+h) = f(y(t), \ldots, y(t-d+1), u(t), \ldots, u(t-q+1)) \tag{4.3}$$

where it is assumed that the samples of the time series $y(t)$ and $u(t)$ are known until time t and the problem is equivalent to a function estimation.

Of course, if exogenous inputs are not considered, expression (4.3) reduces to (4.4)

$$y(t+h) = f(y(t), \ldots, y(t-d+1)). \tag{4.4}$$

In the latter case the problem of estimating $y(t+h)$ is solved by iteratively using expression (4.2). At each step the predicted output is feedback as an argument of the f function. Hence, the f arguments consist of predicted values as opposed to actual observations of the original time series. A prediction iterated for h times returns a h-step-ahead forecasting. The task of forecasting a time series over a long horizon is commonly tackled by iterating one-step-ahead predictors. Despite the popularity that this approach gained in the prediction community, its design is still affected by a number of important unresolved issues, the most important being the accumulation of prediction errors [4].

4.3 EPS Time Series Models

A problem dealing with NAR (and NARX) models is that the vector of regressors $(y(t), y(t-1), \ldots, y(t-d+1))$ of the output variable are the most recent past samples, which very often are correlated each other, i.e., are not really independent variables. To avoid using consecutive regressors of $y(t)$ it is possible to modify the regressor vector as $((y(t), y(t-\tau), \ldots, y(t-(d-1)\tau))$, i.e., the regressors are

time-spaced by τ steps and thus expression (4.1) can be modified as in expression (4.5).

$$y(t+1) = f(y(t), y(t-\tau), \ldots, y(t-(d-1)\tau)). \tag{4.5}$$

The τ parameter is usually referred as the *delay*, which assures that two consecutive regressors of the f function are scarcely correlated and thus almost independent. Expression (4.5) is the so-called embedded phase-space (EPS) representation of dynamical systems [5], which is largely considered in nonlinear modeling of complex systems. Of course expression (4.5) reduces to the traditional NAR form (4.1) when $\tau = 1$. In the framework of EPS models the parameter τ is usually chosen using the criterion of the first minimum of the mutual information (see Sect. 1.7), while the d parameter, referred to as the *embedding dimension* can be chosen in such a way that the fraction of false neighbors (see Sect. 1.11) is negligible.

Of course the MISO map (4.5) can be appropriately extended for multistep prediction according with expression (4.6).

$$\begin{aligned} y(t+h) =& f(y(t), y(t-\tau), \ldots, y(t-(d-1)\tau) \\ & h = 1, 2, ... \end{aligned} \tag{4.6}$$

or with expression (4.7)

$$\begin{aligned} y(t+h) =& f(y(t), y(t-\tau), \ldots, y(t-(d-1)\tau) \\ & u(t), u(t-\tau), \ldots, u(t-(q-1)h)) \\ & h = 1, 2, ... \end{aligned} \tag{4.7}$$

in the presence of exogenous inputs.

4.4 Mapping Approximation

Regardless of which model representation will be chosen by the modeler, among those discussed in the previous section, it is required, in addition to defining the d and τ parameters, approximating the unknown map f. To this purpose, neural networks-based approaches are among the most popular and efficient tools. In this book, two of these kinds of approaches will be considered, namely the neuro-fuzzy and the feedforward neural networks approaches, respectively. One of the main advantages of these approaches is that they allow to approximate nonlinear maps by various kinds of basis function, such as, sigmoidal, gaussian, wavelet, and so on. In particular, the gaussian basis functions seems particularly appropriate for solar radiation time series modeling, due to the gaussian shape of daily solar time series.

4.4.1 The Neuro-Fuzzy Approach

One of the most interesting aspects of the neuro-fuzzy approach is that once the neural network has been trained using automatic learning algorithms, the obtained model can be interpreted in terms of a base of *if ... then* rules. The resulting models can be represented both in linguistic form, or as multidimensional surfaces, whose coordinates are the arguments of the f function. In particular, if the rules are expressed in the so-called Takagi-Sugeno form [6], i.e., with the consequent part expressed as a linear combination of the input mapping, often the model surfaces are iperplanes and thus the rule base can be approximated by simple mathematical expressions. Identification of the model rule base can be obtained in several ways. In particular, the MATLAB® *genfis3.m* function allows to generate a FIS (fuzzy inference system) using the fuzzy c-means clustering algorithm, while the *evalfis.m* function allows the input–output model simulation. Various kinds of functions are allowed to represent the so-called membership functions, such as for instance the Gaussian type, which are particularly appropriate for solar radiation time series modeling. In this book, the combination of the NAR or EPS models and the neuro-fuzzy neural networks for approximating the f map, will be referred, to as the NARNF or EPSNF approach, respectively.

4.4.2 The Feedforward Neural Network Approach

Feedforward networks are probably one of the most popular kinds ofneural networks. They consist of a number of simple artificial neurons, organized in at least three layers. The first layer has a connection from the network input. Each subsequent layer has a connection from the previous layer. The final layer produces the network's output. Such a kind of networks can be used for many kinds of input/output mapping. A celebrated theorem [7] guarantees that a feed-forward network with at least one hidden layer and enough neurons can fit any finite input–output map f, provided that it is continuous. Several training algorithms can be used to learn the network, such as for instance the popular backpropagation or the Levenberg–Marquardt optimization algorithm. In this book, the combination of the NAR or EPS models and the feedforward neural networks for approximating the f map, will be referred, to as the NARNN or EPSNN approach, respectively. In MATLAB®, feedforward neural networks with a predefined number of neurons in the hidden layer can be created using the *feedforwardnet* function, configured for input–output using the *configure* function and trained using the *train* function (see the function List in Appendix A.1).

4.5 Assessing the Model Performances

In order to objectively evaluate the performance of a prediction model several indices can be computed. In this book, two error indices will be considered: the *mae* (mean absolute error) and the *rmse* (root square mean error), since are among the most considered in literature. Such indices are defined as expressed in (4.8) and (4.9), respectively.

$$mae = \frac{1}{n}\sum_{i=1}^{n}|y(i) - \hat{y}(i)| \tag{4.8}$$

$$rmse = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y(i) - \hat{y}(i))^2} \tag{4.9}$$

where n is the number of samples considered to compute the error indices and the symbol $\hat{y}$ indicates the estimated sample.

Often a model is evaluated in comparison with another model, assumed as reference, defining the so-called skill index, S, as in expression (4.10).

$$S = 1 - \frac{rmse_{proposed}}{rmse_{reference}} \tag{4.10}$$

It is trivial to observe that $0 \leq S \leq 1$ if $rmse_{proposed} \leq rmse_{reference}$. In the best case $S = 1$. Negative skill means that the reference model performs better than the proposed one.

4.5.1 Reference Models

Several reference models have been proposed in literature in order to assess the convenience of using some kinds of solar radiation and wind speed prediction models. The main features of a reference model must be its simplicity and/or its ability to predict the deterministic part of the process. According to the criterion of simplicity, the most popular models are probably the P_h and P_{24} models, described below:

- The P_h persistent model is characterized by the simple Eq. (4.11)

$$\hat{y}(t + h) = y(t). \tag{4.11}$$

- The P_{24} persistent model is characterized by Eq. (4.12)

$$\hat{y}(t) = y(t - 24). \tag{4.12}$$

However, the (4.11) model can be trivial in the case of solar radiation, due to the fact that more sophisticated models are able to tacking into account that part of solar radiation which is precisely predictable and traceable. In this case, the so-called clear sky (CSK) model is more appropriate. Some detail about the CSK model, implemented as the *pvl_clearsky_ineichen* function, will be given in Chap. 5.

4.6 Conclusions

A huge number of techniques have been proposed in literature to model solar radiation and wind speed time series and thus it is almost impossible to be exhaustive dealing with this subject. For this reason, the description was limited to the NAR and EPS approaches, which, in conjunction with neural networks-based approaches to identify the unknown mapping function, are considered the most appropriate for the purposes of this book.

References

1. F. Takens, Detecting strange attractors in turbulence, in *Dynamical Systems and Turbulence*, ed. by D.A. Rand, L.-S. Young. Lecture Notes in Mathematics, vol. 898 (Springer-Verlag, 1981), pp. 366–381
2. L. Ljung, *System Identification—Theory for the User*, 2nd edn. (Prentice-Hall, Upper Saddle River, N J, 1999)
3. S.A. Billings, in *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains* (Wiley, 2013). ISBN: 978-1-1199-4359-4
4. Q. Zhang, L. Ljung, *Multiple steps prediction with nonlinear arx models, in Proceedings NOLCOS*, *IFAC Symposium on Nonlinear Control Systems*, Germany, Stuttgart, Sept 2004
5. H. Kantz, T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, 1997)
6. T. Takagi, M. Sugeno, Fuzzy identification of systems and its applications to modeling and control. IEEE Trans. Syst. Man, Cybern. **1**, 116–132 (1985)
7. G. Cybenko, Approximations by superpositions of sigmoidal functions. Math. Control Signals Systems **2**(4), 303–314 (1989)

Chapter 5
Modeling Hourly Average Solar Radiation Time Series

Abstract This chapter deals with the problem of short-term prediction of hourly average solar radiation time series, recorded at ground level, by using embedding phase-space (EPS) models. Two different neural approaches have been considered to identify the nonlinear map underlying the identification problem, namely the neuro-fuzzy (NF) approach and the feedforward neural network (NN) approach. Performances are evaluated in terms of *mae*, *rmse* and skill index, in comparison with two popular reference models, namely the clear sky model and the P_{24} persistent model.

5.1 Introduction

Several approaches for short-term prediction of solar radiation time series have been proposed in literature, such as ARMA (auto regressive moving average) and TDNN (time delay neural networks) [1, 2], recurrent neural networks [3], ANFIS (adaptive neuro-fuzzy inference systems) and ANN (artificial neural networks) [4, 5], particle swarm optimization, and evolutionary algorithm, using recurrent neural networks [6], ARMA-GARCH (generalized autoregressive with conditional Heteroskedasticity) models [7], Bayesian statistical models [8], fuzzy with genetic algorithms models by [9] and machine learning approaches [10]. Statistical approaches based on decomposition of the original time series models have been studied by [11–13].

In this chapter, it is proposed to use the nonlinear parametric approach referred to as the embedding phase-space (EPS) models, described in the previous Sect. 4, to model hourly average solar radiation time series. The parameters of such a kinds of models will be estimated by using ANFIS and feedforward neural network approaches. In order to objectively asses the performances of prediction models the *mae* (mean absolute error) and the *rmse* (root square mean error), expressed by (4.8) and (4.9), respectively will be computed. Furthermore, the model performances will be evaluated also in terms of skill index (see expression (4.10)) by inter-comparing the EPS models with the so-called clear sky model (CSK). In particular, the Ineichen and Perez CSK model for global horizontal irradiance as presented in [15, 16] will be considered. The MATLAB® code that implement this model was obtained from the

L. Fortuna et al., *Nonlinear Modeling of Solar Radiation and Wind Speed Time Series*, SpringerBriefs in Energy, DOI 10.1007/978-3-319-38764-2_5

Sandia National Labs PV Modeling Collaborative (PVMC) platform (see Appendix A.2 for details).

5.2 Modeling Results

Model performances are described in Sects. 5.2.1 and 5.2.2 for EPSNF and EPSNN models, respectively. Afterwards a direct comparison between these two approaches is described in Sect. 5.2.3. In the modeling process nighttime data was not excluded in order to avoid of dealing with the different duration of the solar day through the year. Model performances was evaluated for intra-day forecasting, i.e., assuming $h \in [1, 24]$. For modeling purposes, three years of data was considered for each recording station, from 2003 to 2005. In particular, two years of data, from 2003 to 2004, was considered to identify the model parameters while data recorded during 2005 was reserved for testing the model. Parameters of the EPSNF models were identified by using the *genfis3.m* MATLAB® function, choosing three clusters for each argument of the f map. To identify the EPSNN models the *feedforwardnet.m* MATLAB® function was considered, setting the number of neurons in the hidden layer to 20.

5.2.1 Performances of the EPSNF Approach

For simplicity, results will be described by considering a particular recording station, such as, for instance, the New York Central PRK (station ID725033) solar station. Afterwards, results obtained for the others considered stations will be shown. The performances in terms of *mae* and *rmse* are reported in Fig. 5.1 for different embedding dimensions. The horizontal lines in Fig. 5.1a, b represent the level of *mae* and *rmse* of the CSK model, respectively, which is independent of the prediction horizon. It is possible to see that the EPSNF models outperform the clear sky model provided that the embedding dimension is appropriately chosen ($d \geq 8$). For all trials presented in this section the delay parameter was set to $\tau = 3$ (i.e., half of the value obtained by using the criterion of the first minimum of the mutual information obtained in Sect. 2.2.2). One of the most interesting aspects of using the NF approach to identify the f map in expression (4.6), is that the identified models can be easily interpreted. Indeed, it is possible to show that by using the Takagi–Sugeno form to represent the rules, it is possible to obtain that f assumes the form of iper-planes and thus expression (4.6) simplifies as (5.1)

$$GHI(t+h) = \sum_{k=1}^{d} a_k GHI(t-(k-1)\tau) + a_{d+1} \tag{5.1}$$

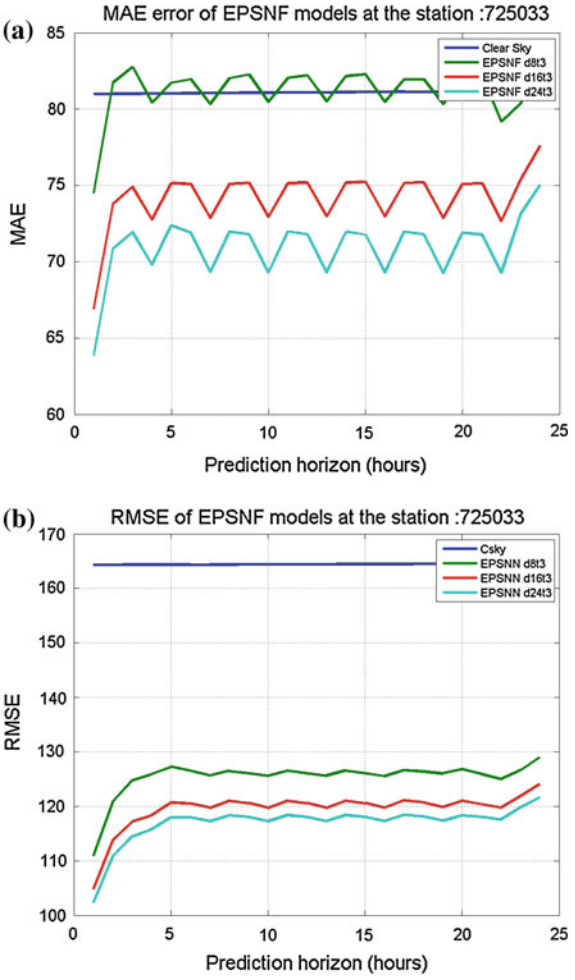

Fig. 5.1 Performances of the EPSNF models featured by $d \in [8, 16, 24]$ and $\tau = 3$ **a** mae, **b** rmse

For instance the EPSNF model at the station ID 72503, featured by $d = 8, \tau = 3$ and $h = 5$, can be represented as in (5.2).

$$\begin{aligned} GHI(t+5) = &- 0.0690GHI(t) + 0.5164GHI(t-3) + \\ &0.3487GHI(t-6) - 0.0653GHI(t-9) + \\ &- 0.0905GHI(t-12) + 0.0271GHI(t-15) + \\ &- 0.1229GHI(t-18)GHI(t-21) + 0.0315 \\ &+ 65.2767 \end{aligned} \tag{5.2}$$

and its time behavior, in comparison with the true time series, is shown in Fig. 5.2. It is possible to see that despite its extreme simplicity, since it is an approximated linear

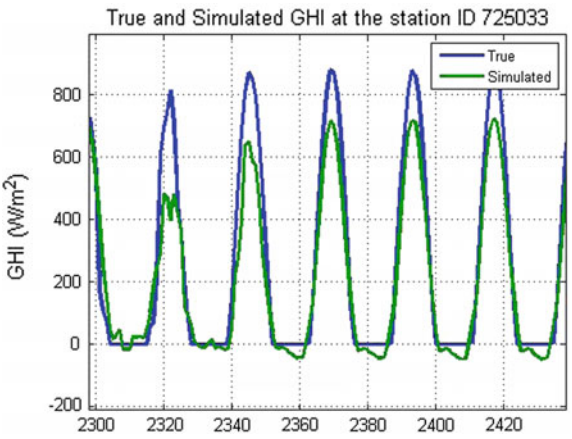

Fig. 5.2 Time behavior of the EPSNF model (5.1) obtained for the station ID 725033

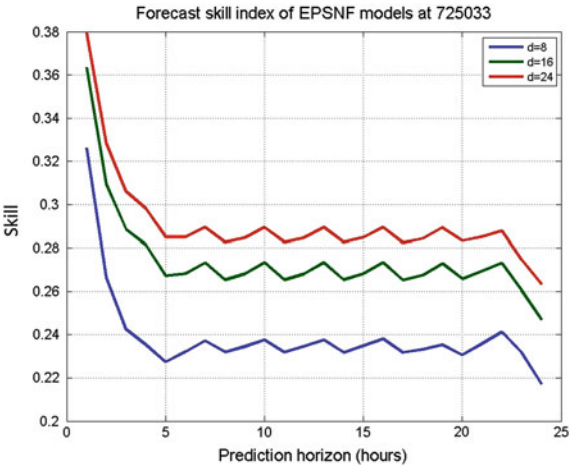

Fig. 5.3 Skill index of EPSNF models for three different embedding dimensions, versus the prediction horizon, for the station ID 725033

model, it is able to reproduce the general behavior of the true time series. The skill index of EPSNF models, for three different embedding dimensions ($d = 8$, $d = 16$ and $d = 24$), and $\tau = 3$, versus the prediction horizon, for the station ID725033, is shown in Fig. 5.3. It is possible to see that:

- The skill index for these kinds of models depends on the embedding dimension d: higher d is better.
- The skill index decreases with the prediction horizon, reaching an asymptotic value in correspondence of $h = 5$. This result agrees with the finding reported in the previous Sect. 2.2.2, where it was shown that autocorrelation and mutual information of hourly average time series reaches also an asymptotic value after 5 lags.
- The EPSNF models outperform the CSK model. In the best case reported in Fig. 5.3, i.e., for $d = 24$, $S \in [0.28, 0.38]$.

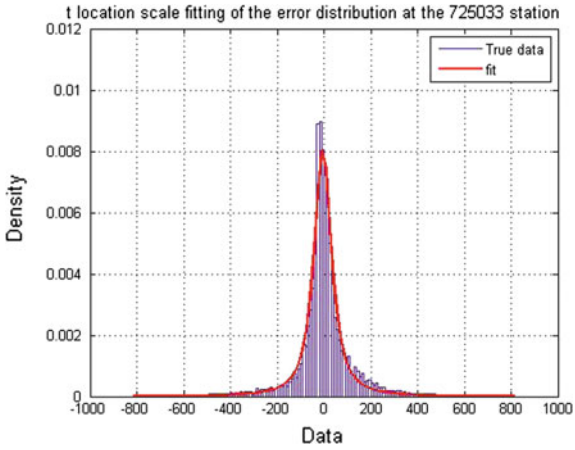

Fig. 5.4 t-Location scale distribution density of the hourly average error generated by the EPSNF model ($h = 5$, $d = 24$, $\tau = 3$) at station ID 725033 during the test year

5.2.1.1 Analysis or the Residual

In order to better evaluate the model performances, the empirical probability distribution density of the model residual was computed and then fitted by using known continuous probability distribution density functions (*pdf*). To this purpose, it was found that the t-Location *pdf* was appropriate to fit the residual of the considered EPSNF prediction model, as shown in Fig. 5.4. Indeed, the t-location-scale is useful for modeling data distributions with heavier tails, i.e., more prone to outliers, than the normal distribution. Such a *pdf* is characterized by three parameters, referred to as μ, the location parameter, σ the scale parameter, and ν, the shape parameter. While $-\infty < \mu < \infty$, the two remaining parameters are constrained to be $\sigma > 0$ and $\nu > 0$. It can be demonstrated that the t-location-scale distribution approaches the normal distribution as ν approaches infinity, and that smaller values of ν yield heavier tails. Furthermore, in the hypothesis that $\nu > 1$ the mean of the t-location-scale distribution is μ, i.e., it coincides with the location parameter; otherwise the mean is not defined. Similarly, in the hypothesis that $\nu > 2$ the variance of the t-location distribution can be computed as $var = \sigma^2\nu/(2-\nu)$, otherwise it is not defined. In the fitting shown in Fig. 5.4 the three parameters assumes the following values: $\mu = -4.97$, $\sigma = 41.81$, $\nu = 1.39$. This implies that μ, the location parameter, represents also the mean of the error, which is slightly negative, i.e., this model tends to under estimate the true GHI time series.

5.2.1.2 Performances of EPSNF Models for All Considered Stations

Performances, in terms of skill index, of EPSNF models obtained for all recording stations considered in this work, setting $d = 24$ and $\tau = 3$, are shown in Fig. 5.5. As expected, the skill index behaves quite similarly to that described for the station

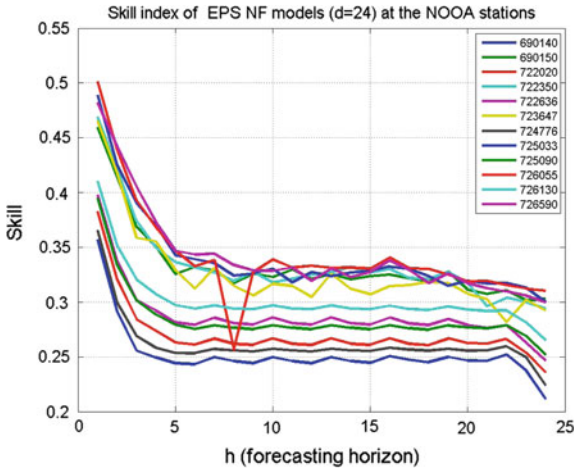

Fig. 5.5 Skill index for all considered stations during the test year

Table 5.1 Parameters of the tlocation *pdf* for h = 5, fitted for the 12 considered stations, by using EPSNF models featured by $d = 24$ and $\tau = 3$

Station	μ	σ	ν
690140	3.17	29.49	1.28
690150	6.62	26.29	1.46
722020	1.28	41.48	1.36
722350	0	41.92	1.38
722636	3.79	33.58	1.28
723647	5	33.98	1.34
724776	5.36	29.98	1.18
725033	−4.97	41.81	1.39
725090	−6.1	41.66	1.37
726055	−7.64	42.09	1.37
726130	−10.21	34.44	1.17
726590	−4.14	32.98	1.22

ID 725033. However, it is possible to see a variability from station to station of the level of skill, which could be attributed to the local meteo-climate conditions. As concerning the characterization of model residues for a prediction horizon set to $h = 5$, the t-location scale *pdf*, fitted for each of the considered stations, are shown in the Table 5.1. As it is possible to see, since for all stations is $1 < \nu < 2$, the mean of the t-location *pdf* is defined and it is represented by the location parameter μ.

5.2.2 *Performances of the EPSNN Approach*

The aim of this section is to describe the performances of EPS models of the form (4.6) where the f map will be approximated by using the feedforward neural network approach. The performances, in terms of *MAE* and *RMSE*, of EPSNN model featured by $d = [8, 16, 24]$ and $\tau = 3$, computed for the station ID 725033, are reported in Fig. 5.6. It is possible to see that both the level of *MAE* and *RMSE* for the EPSNN model are significantly below the corresponding values assumed by the reference

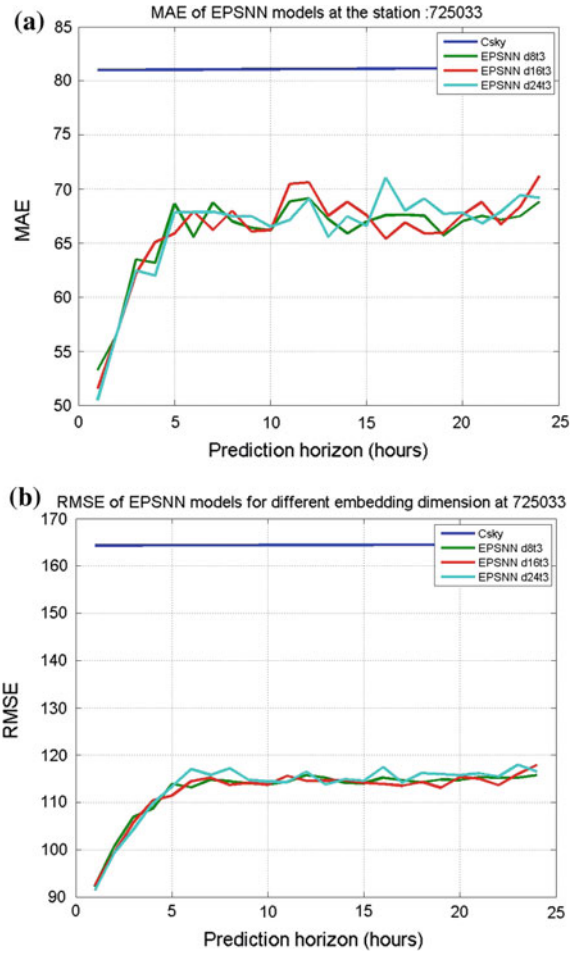

Fig. 5.6 Performances of the EPSNN model featured by $d = 8$ and $\tau = 3$ for the station ID725033. **a** mae, **b** RMSE

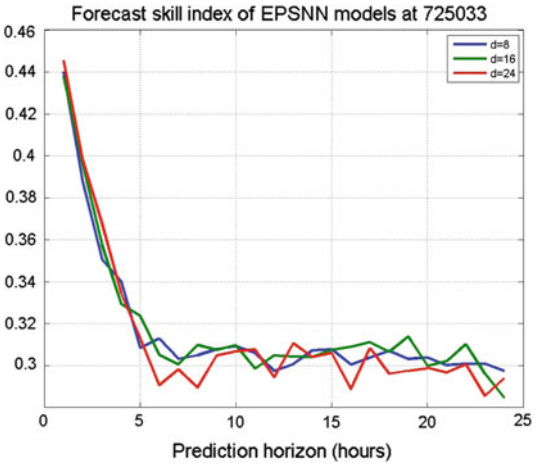

Fig. 5.7 Skill index of EPSNN models with embedding dimension $d = 8$, $d = 16$ and $d = 24$, respectively, obtained for the station ID 725033

CSK model. The skill index of this model is shown in Fig. 5.7. It is possible to see that:

- The skill index of EPSNN models featured by $d = 8$, $d = 16$, and $d = 24$ are quite similar.
- The skill index decreases with the prediction horizon, reaching an asymptotic value in correspondence of $h = 5$. This result agrees with the finding reported in the previous Sect. 2.2.2, where it was shown that autocorrelation and mutual information of hourly average time series reaches also an asymptotic value after lag $= 5$.
- The EPSNN models outperform the CSK model since $S \in [0.25, 0.45]$. It is trivial to observe that the higher skill values correspond to the shorter prediction horizons.

5.2.2.1 Analysis or the Residual

In order to characterize the error distribution, the t-Location *pdf* was considered to fit the residual of the considered EPSNN prediction model described above, as shown in Fig. 5.8. For the station ID 725033 the parameters of the best t-location *pdf* assume the values $\mu = 1.8$, $\sigma = 23.82$ and $\nu = 0.91$, as reported in Table 5.2.

5.2.2.2 Model Performances at All Considered Stations

Performances, in terms of skill index, of EPSNN models obtained for all recording stations considered in this work, setting $d = 24$, and $\tau = 3$, are shown in Fig. 5.9. As expected, the skill index behaves quite similarly to that described for the station ID 725033. However, it is possible to see a significant variability from station to

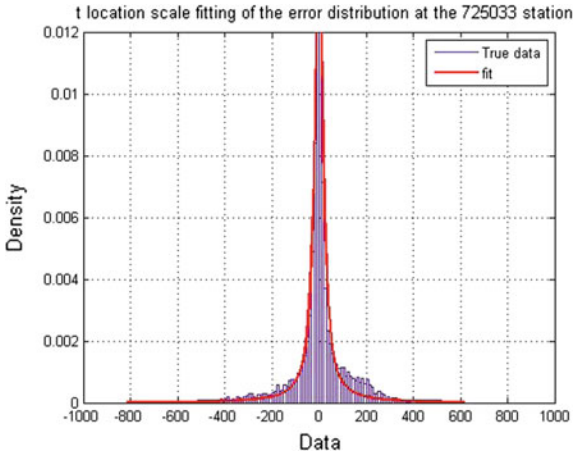

Fig. 5.8 t-Location scale distribution density of the hourly average error generated by the EPSNN model ($h = 5$, $d = 24$ and $\tau = 3$) at the station ID 725033 during the test year

Table 5.2 Parameters of the tlocation *pdf* fitted at the 12 considered stations for h = 5 by using EPSNN featured by $d = 24$ and $\tau = 3$

Station	μ	σ	ν
690140	0.3	14.13	0.82
690150	−0.55	14	1.07
722020	0.67	26.01	0.99
722350	1.25	29.38	1.06
722636	0.79	16.57	0.83
723647	1.95	17.38	0.91
724776	0.49	18.62	0.91
725033	1.8	23.82	0.91
725090	0.48	21.84	0.87
726055	0.69	23.28	0.9
726130	−5.04	18.93	0.79
726590	1.29	20.21	0.86

station of the level of skill, which, as observed in the previous Sect. 5.2.2, should be attributed to the local meteo-climate conditions. As concerning the characterization of model residues for a prediction horizon set to $h = 5$, the t-location scale *pdf*, fitted for each of the considered stations are shown in Table 5.2.

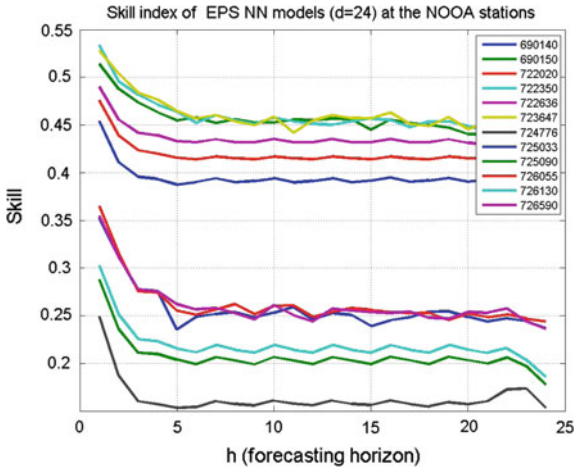

Fig. 5.9 Skill index of EPSNN models for all the considered stations

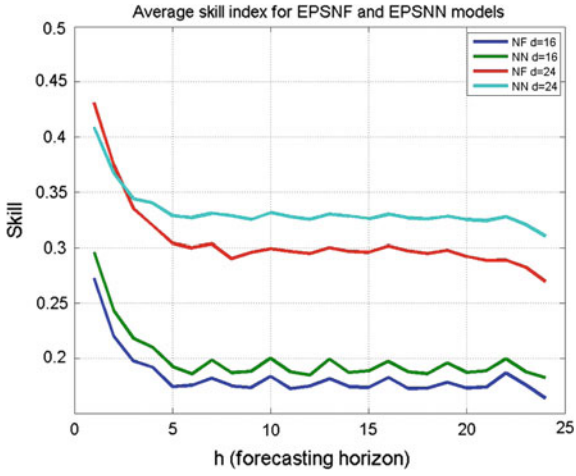

Fig. 5.10 Average skill of EPSNF and EPSNN models for $d = 16$ and $d = 24$ and $\tau = 3$

5.2.3 A Direct Comparison Between EPSNF and EPSNN

The skill index of EPSNF and EPSNN, averaged over the 12 stations considered in this work during the test year, for the whole explored forecasting horizons, and for two embedding dimensions ($d = 16$ and $d = 24$), are shown in Fig. 5.10. This Figure shows that:

- Models characterized by $d = 24$ behave, on average, better than models featured by $d = 12$.
- EPSNN models outperform on average EPSNF models with similar d and τ.
- The skill index S decays, on average, from about 0.43, obtained for $h = 1$ to about 0.30 for $h = 5$. For higher values of h it holds the level reached for $h = 5$. This

result agrees with the autocorrelation analysis, which estimates the correlation time τ_c of hourly average solar radiation time series in about five lags. Nevertheless, there is some convenience on using the proposed models for $5 < h \leq 24$ with respect to the clear sky model.

5.2.4 *Average Skill Index Considering the P_{24} Reference Model*

Another reference model often considered in the literature, dealing with solar radiation forecasting models, is the so-called 24 h persistence, i.e., a model expressed by (5.3), i.e., a model which assumes as that the solar radiation at the generic hour t is the same as that recorded yesterday at the same hour.

$$GHI(t) = GHI(t - 24). \tag{5.3}$$

Using such a model as the reference model, the skill of EPSNN models having $d = 24$ and $\tau = 3$, computed for each individual station is shown in Fig. 5.11. A direct comparison between EPSNF and EPSNN models for two different embedding dimensions ($d = 16$ and $d = 24$), obtained comparing the skill averaged over the 12 stations is shown in Fig. 5.12. It is possible to observe that EPSNN models perform better than EPSNF models also considering the P_{24} as a reference model. Furthermore, although the average skill of the EPS models compared to the P_{24} model is slightly lower than that computed by using the CSK reference model, the EPS models still significantly outperform the reference model.

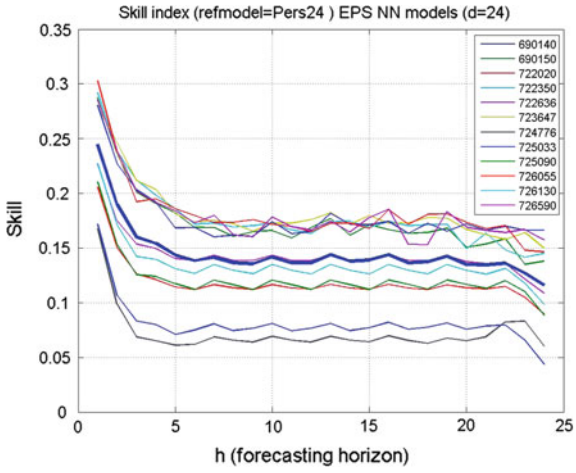

Fig. 5.11 Skill of EPSNN models with $d = 24$ and $\tau = 3$ against the 24 h persistence reference model. The thick curve represents the skill obtained averaging over the 12 considered stations

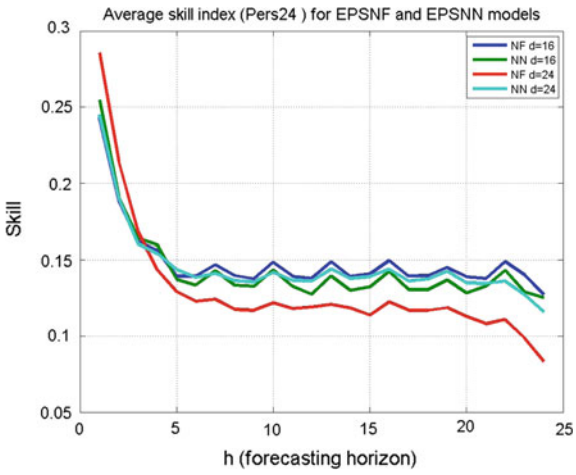

Fig. 5.12 Average skill of EPSNF and EPSNN models for two different embedding dimension ($d = 16$ and $d = 24$) and $\tau = 3$, against the 24 h persistence reference model

5.3 Conclusions

In this work, the problem of analyzing and modeling solar radiation time series was addressed. The studied prediction models are based on information that can be gathered from time series recorded at the site of interest only, thus excluding in the modeling process any other information, including the fact that the processes involved are spatially distributed. This radical choice of field was carried out since the main aim of the studied prediction models, is that of being agile, in contrast with the prediction models of type NWP (Numerical Weather Prediction), which are more complex and thus could be not appropriate for several reasons. Results described show that the proposed EPS based approaches, provided that the structural parameters τ and d are appropriately chosen, provides, on average, a skill index, evaluated in comparison with the clear sky model, in the range $S \in [0.30, 0.43]$ for prediction horizons in the range $h \in [1, 24]$. Of course, the best performance, in terms of skill index are obtained for shorter horizons in the range $h \in [1, 5]$. Results have pointed out that EPSNN models are usually more accurate then EPSNF models. However, these latter kinds of models could be preferred when a simple input-output representation is needed. Another aspect of the analysis carried out is the dependence of model performance from on the considered recording site which could be attributed to the local meteo-climate conditions. It would be interesting to investigate the causes of this variability, but it was outside the scope of this book.

References

1. A.M.A. Baig, P. Achter, Prediction of hourly solar radiation using a novel hybrid model of arma and tdnn. Solar Energy **1**, 119–123 (1991)
2. J. Wu, C.K. Chan, Prediction of hourly solar radiation using a novel hybrid model of arma and tdnn. Solar Energy **85**, 808–817 (2011)
3. Z. Nian, P. Behera, Solar radiation prediction based on recurrent neural networks trained by levenberg-marquardt backpropagation learning algorithm, in *Innovative Smart Grid Technologies (ISGT)*, 2012 IEEE PES, pp. 1–7 (2012). doi:10.1109/ISGT.2012.6175757
4. A. Yona, T. Senjyu, T. Funabashi, C. Kim, Determination method of insolation prediction with fuzzy and applying neural network for long-term ahead pv power output correction. IEEE Trans. Sustain. Energy **4**, 527–533 (2013)
5. J. Piri, O. Kisi, Modelling solar radiation reached to the earth using anfis, nn-arx, and empirical models (case studies: Zahedan and bojnurd stations). J. Atmos. Sol. Terr. Phys. **123**, 39–47 (2015)
6. Z. Nian, P. Behera, C. Williams, Solar radiation prediction based on particle swarm optimization and evolutionary algorithm using recurrent neural networks, in *IEEE International Systems Conference (SysCon)* (2013)
7. H. Sun, D. Yan, N. Zhao, J. Zhou, Empirical investigation on modeling solar radiation series with armagarch models. Energy Convers. Manag. **92**, 385–395 (2015)
8. P. Lauret, J. Boland, B. Ridley, Bayesian statistical analysis applied to solar radiation modelling. Renew. Energy **49**, 124–127 (2013)
9. O. Kisi, Modeling solar radiation of mediterranean region in turkey by using fuzzy genetic approach. Energy Convers. Manag. **64**, 429–436 (2014)
10. S. Navin, S. Pranshu, I. David, S. Prashant, Predicting solar generation from weather forecasts using machine learning, in *IEEE International Conference on Smart Grid Communications (Smart Grid Comm)*, Brussels, Belgium
11. V. Prema, K.U. Rao, Development of statistical time series models for solar power prediction. Renew. Energy **83**, 100–109 (2015)
12. W. Yao, Z. Li, T. Xiu, Y. Lu, X. Li, New decomposition models to estimate hourly global solar radiation from the daily value. Solar Energy **120**, 87–99 (2015)
13. P. Alvanitopoulos, I. Andreadis, N. Georgoulas, M. Zervakis, N. Nikolaidis, Solar radiation prediction model based on empirical mode decomposition, in *2014 IEEE International Conference on Imaging Systems and Techniques (IST)*, pp. 161–166 (2014)
14. L. Mazorra Aguiar, B. Pereira, M. David, F. Diaz, P. Lauret, Use of satellite data to improve solar radiation forecasting with Bayesian Artificial Neural Networks. Solar Energy **122**, 1309–1324 (2015)
15. P. Ineichen, R. Perez, A New airmass independent formulation for the Linke turbidity coefficient. Phys. A **73**, 151–157 (2002)
16. R. Perez, A new operational model for satellite-derived irradiances-description and validation. Solar Energy **73**, 207–317 (2002)

Chapter 6
Modeling Hourly Average Wind Speed Time Series

Abstract In this Chapter results obtained by applying the modeling approaches described in Chap. 4 to the WWR data set of hourly average time series are reported. For all modeling trials, data recorded during 2004 and 2005 was considered to identify the model parameters while the 2006 was reserved to test the model. Performances have been evaluated in terms of *mae*, *rmse* and skill index, in comparison with the P_h persistent model.

6.1 Introduction

Wind speed is an important renewable energy source and a lot of efforts have been devoted in literature to study effective techniques to predict wind speed time series. A review of the young history of methods for short term prediction was given in [1]. Another good review can be found in [2]. ARMA models have been considered by [3], while Autoregressive Integrated Moving Average (ARIMA) and the artificial neural network (ANN) methods have been proposed by [4]. Neuro-fuzzy based techniques were considered by [5, 6]. Soft computing models were considered by [7] and empirical mode decomposition (EMD), chaotic theory, and grey theory were proposed by [8]. In this Chapter it is proposed the use of NAR and EPS models, identified by using both the NF and FF approaches.

6.2 Considerations on the Choice of Model Parameters

Following the scheme considered in the previous chapter for the solar radiation time series, we begin showing the performances of EPSNF models featured by $\tau = 3$ and three different embedding dimension $d \in [8, 16, 24]$ at one station (the ID2257). Figure 6.1a, b show, respectively, the *mae* and the *rmse* of the considered models in comparison with the persistent model, while Fig. 6.1c shows the skill index.

L. Fortuna et al., *Nonlinear Modeling of Solar Radiation and Wind Speed Time Series*, SpringerBriefs in Energy, DOI 10.1007/978-3-319-38764-2_6

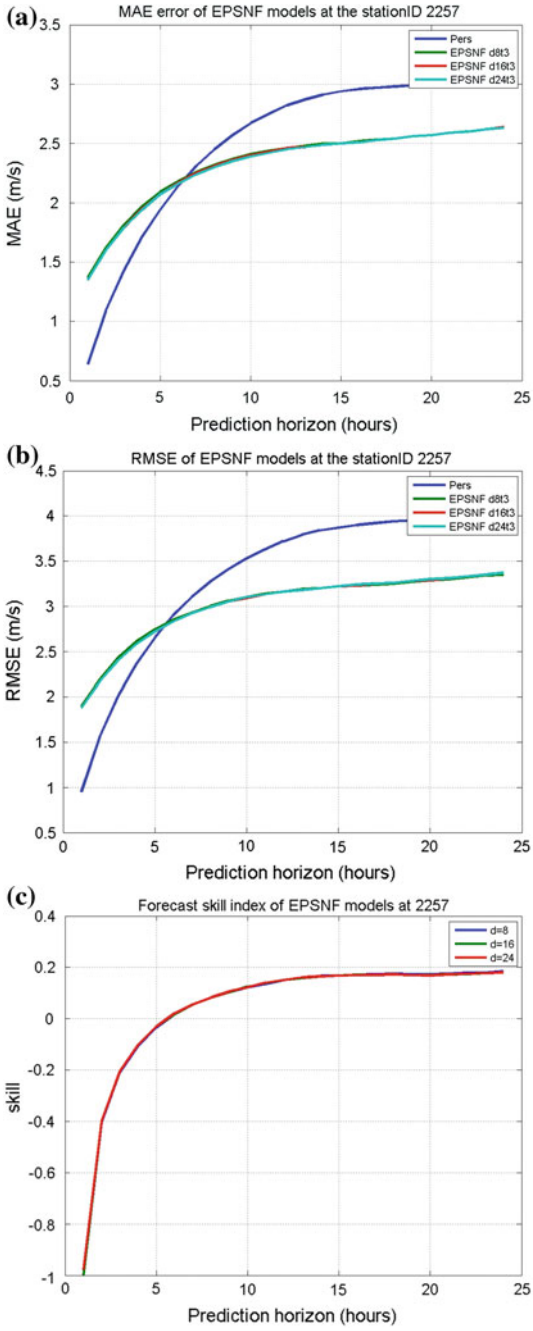

Fig. 6.1 Forecasting errors of EPSNF models with $\tau = 3$ and $d \in [8, 16, 24]$ at the station ID2257. **a** mae. **b** *rmse*. **c** Skill index

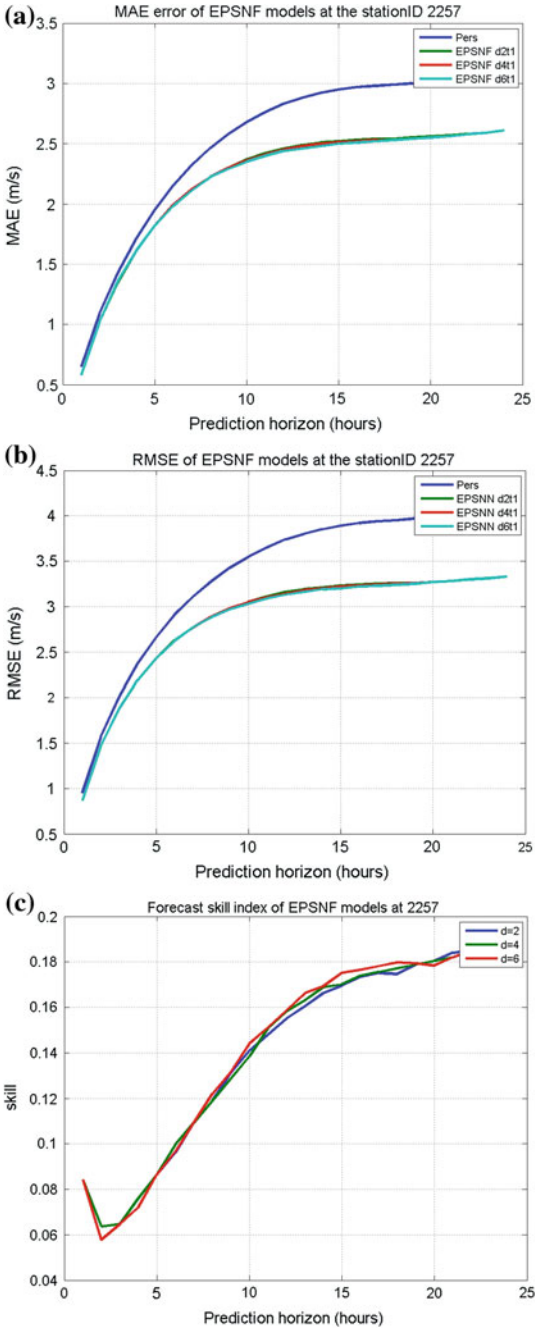

Fig. 6.2 Forecasting errors of EPSNF models with $\tau = 1$ and $d \in [2, 4, 6]$ at the station ID2257. **a** *mae*. **b** *rmse*. **c** Skill index

As it is possible to see, the error, both in terms of *mae* and *rmse* grows rapidly for low h, say $1 < h < 6$. For higher values of h, the error continues to grow, but with a lower rate. This behavior is the opposite of what happens to the mutual information of hourly averages of wind speed, already pointed out in Sect. 3.4, where it was seen that these functions decrease rapidly up to lag 6 and then reaches a low steady-state value. Furthermore, it is possible to realize, for the chosen model parameters, the persistent model outperforms the corresponding EPSNF models for prediction horizon in the range $h \in [1, 5]$, while for $h > 5$ the contrary occurs.

A different choice of τ and d, for instance $\tau = 1$ and $d \in [2, 4, 6]$, allows to improve the performance of the EPS models, with respect to the persistent model, as shown in Fig. 6.2, but in essence EPS models perform almost as the persistent model in the range $h \in [1, 5]$ which is the most important for applications, since for higher h, the error is unacceptably high. This means that low order models work better than high-order models for the prediction of the hourly averages of wind speed. In other terms, models with several regressors, tend to averaging among several samples, losing the ability to forecast wind speed time series. Indeed, as shown in Chap. 3, this kind of time series are irregularly fluctuating at any time scale. Furthermore, several trials, carried out assuming $\tau = 1$, which in practice means that the EPS models reduces to NAR models, have experimentally demonstrate that:

- No significant differences were observed by taking values of $d \in [2, 4, 6]$; thus, there is no reason to consider more than 2 regressors to predict hourly average of wind speed.
- There is not significant difference between EPSNF, EPSNN and the simple persistent model.

6.3 Performances for All the Considered Stations

The *mae* and *rmse* obtained for the twelve considered stations are reported in Fig. 6.3a, b, respectively. It is possible to see that the behavior, in terms of *mae* and *rmse*, already observed for the station ID2257, is shared by all stations of the considered dataset. In the Figures the average *mae* and *rmse* curves are indicated by the thick curves. The skill index, which summarize the performance of the NAR model (see (4.4)) against the persistent model (see (4.11)), is reported in Fig. 6.4. It can be observed that for short prediction horizons, $h \in [1, 3]$, the skill index is lower than 0.1 and therefore negligible. By using the *mae* and *rmse* averaged over the 12 considered stations, we got the expressions (6.1) and (6.2) that allow to synthesize the average accuracy obtained by the studied NAR models.

$$mae_{ave}(h) = 3.317(1 - e^{-0.2502h}) + 0.3433, h = 1, 2, ..., 24 \tag{6.1}$$

$$rmse_{ave}(h) = 4.028(1 - e^{-0.2581h}) + 0.7239, h = 1, 2, ..., 24. \tag{6.2}$$

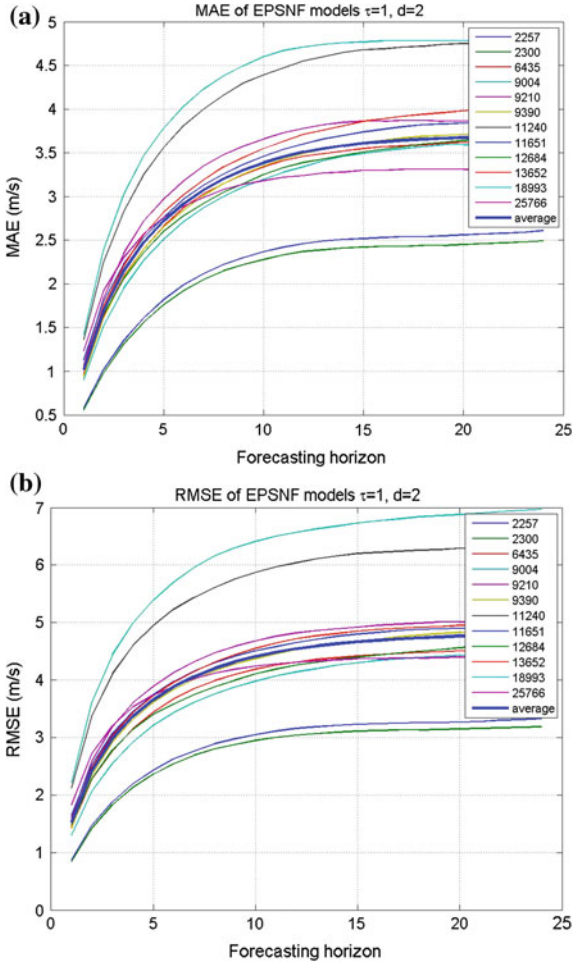

Fig. 6.3 Forecasting errors of EPSNF models with $\tau = 1$ and $d = 2$ at 12 different stations. **a** *mae*. **b** *rmse*

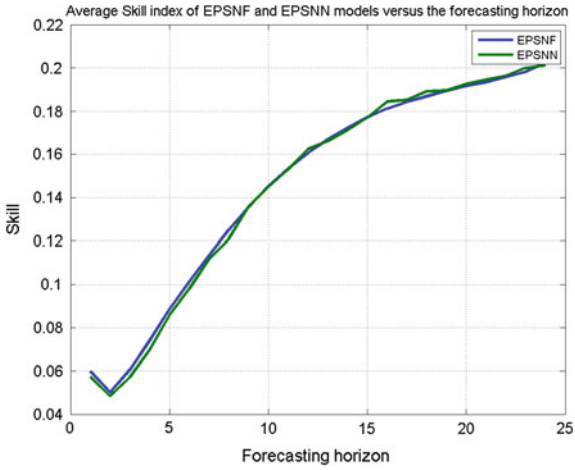

Fig. 6.4 Skill, averaged over 12 station of NAR model against the persistent model

6.4 Conclusions

Autoregressive models described in this chapter are appropriate to predict the hourly average wind speed only for short time intervals, say $1 < h \leq 3$. Nevertheless, for some application, 3 hours ahead predictions can be useful for plant managers to dispatch conventional generators in order to satisfy the electricity demand from the users. However, it must be observed that there is not a decisive convenience in using EPS or NAR models, since they exhibit performances which are only slightly better than the simple persistent model.

References

1. A. Costa, A. Crespo, J. Navarro, G. Lizcano, H. Madsend, E. Feitosa, A review on the young history of the wind power short-term prediction. Renew. Sustain. Energy Rev. **12**, 1725–1744 (2008)
2. G. Giebel, R. Brownsword, G. Kariniotakis, M. Denhard, C. Draxl, The state of the art in short-term prediction of wind power—a literature overview, Project ANEMOS Deliverable Report D1.1, 2003 (2011), http://www.anemos-plus.eu/, 1–110
3. J.L. Torres, A. Garcia, M.D. Blas, A. DeFrancisco, Forecast of hourly average wind speed with ARMA models in Navarre (Spain). Sol. Energy **79**, 65–77 (2005)
4. E. Cadenas, W. Rivera, Wind speed forecasting in three different regions of mexico, using a hybrid ARIMA—ANN model. Renew. Energy **35**, 2732–2738 (2010)
5. M. Mohandes, S. Rehman, S. Rahman, Estimation of wind speed profile using adaptive neuro-fuzzy inference system (ANFIS). Appl. Energy **88**, 4024–4032 (2011)

6. S. Shamshirband, D. Petkovic, N. Anuar, M. Kiah, S. Akib, A. Gani, Z. Cojbaic, V. Nikolic, Sensorless estimation of wind speed by adaptive neuro-fuzzy methodology. Electr. Power Energy Syst. **62**, 490–495 (2014)
7. A.U. Haque, P. Mandal, M.E. Kaye, J. Meng, L. Chang, T. Senjyu, A new strategy for predicting short-term windspeed using soft computing models. Renew. Sustain. Energy Rev. **16**, 4533–4573 (2012)
8. X. An, D. Jiang, M. Zhao, C. Liu, Short-term prediction of wind power using EMD and chaotic theory. Commun. Nonlinear Sci. Numer. Simul. **17**, 1036–1042 (2012)

Chapter 7
Clustering Daily Solar Radiation Time Series

Abstract This chapter deals with the clustering of solar radiation daily patterns. The problem is tackled using a feature-based approach, in order to deal with patterns in a low-dimensional space. To this purpose, an original pair of indices is introduced, referred to as the area ratio A_r and the normalized GPH hurst exponent GPH_r. It is shown that using these features, solar radiation daily patterns can be classified into to 4 classes referred to as completely cloudy, partially cloudy, partially clear sky, and clear sky. Furthermore, it is shown how some useful statistical properties, such as the class weight and the persistence of patterns in a class, can be estimated.

7.1 Two Features of Solar Radiation Time Series

In this section we introduce two features, referred to as A_r and GPH_r, representing the normalized amount of solar energy per day reaching a horizontal surface and the normalized GPH Hurst exponent. While the former feature is useful for representing the daily amount of energy, the latter express its degree of fluctuation.

7.1.1 The Area Ratio A_r Index

The A_r index is formally represented by expression (7.1)

$$A_r(d) = \frac{A_{pat}(d)}{A_{csk}(d)} \tag{7.1}$$

where $A_{pat}(d)$ and $A_{csk}(d)$ represent the area under the true and the clear sky daily solar radiation patterns, respectively, at the generic Julian day d. The A_{csk} term can be computed referring to one of several existing clear sky models. In this book we have considered the Ineichen and Perez CSK model for global horizontal irradiance, as presented in [1, 2]. The MATLAB® code to implement this model is the *pvl_clearsky_ineichen* function (see Appendix A.1). Of course the A_r index is always positive but can be greater or less than 1. For example, in a day featured by favorable

L. Fortuna et al., *Nonlinear Modeling of Solar Radiation and Wind Speed Time Series*, SpringerBriefs in Energy, DOI 10.1007/978-3-319-38764-2_7

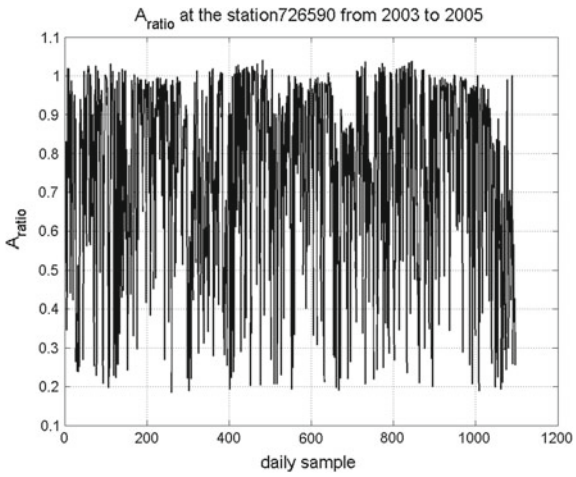

Fig. 7.1 A_r index computed at the station ID726590 from 2003 to 2005

weather conditions (e.g., absence of cloud cover and good atmospheric transmittance) A_r will be greater than 1; conversely under thick cloud cover and adverse propagation conditions it may be significantly less than 1. As an example, the $A_r(d)$ index, computed during 3 years for the station ID726590, is reported in Fig. 7.1.

7.1.2 The GPH_r Index

The GPH_r index is formally represented by expression (7.2)

$$GPH_r(d) = \frac{GPH_{pat}(d)}{GPH_{csk}(d)} \tag{7.2}$$

where $GPH_{pat}(d)$ and $GPH_{csk}(d)$ represent the Hurst exponent computed using the Geweke-Porter-Hudak (GPH) algorithm of the true and clear sky solar radiation pattern, respectively, at the generic Julian day d. The GPH algorithm, first described by [3] was computed using the *gph* function listed in Appendix A.1. The reason for using this algorithm for computing the hurst exponent, instead of R/S and DFA algorithms, is that it is considered more appropriate when patterns are represented by a small number of samples. It is based on the slope of the spectral density function. An example of GPH_r index, computed at the station ID726590 during 3 years is shown in Fig. 7.2. The extremely scattered behavior of these features was pointed out by computing the mutual information of the individual A_r and GPH_r, as shown in Fig. 7.4. It is possible to see that after lag 1, i.e., 1 day, the mutual information reaches the lower values. This result says that 1-day ahead prediction of the class is extremely difficult using autoregressive models. Therefore, clustering approaches can be useful to extract some statistical information (Fig. 7.3).

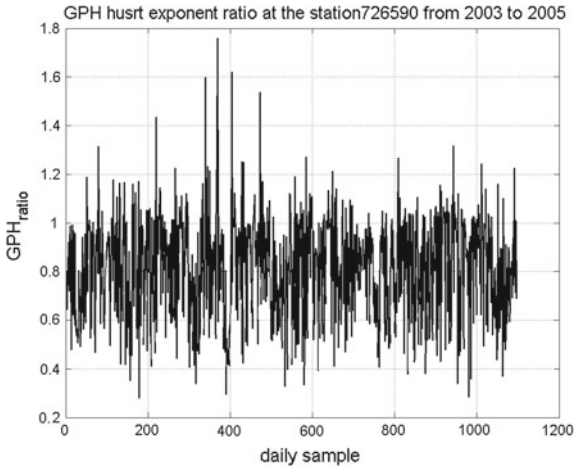

Fig. 7.2 GPH_r index computed at the station ID726590 from 2003 to 2005

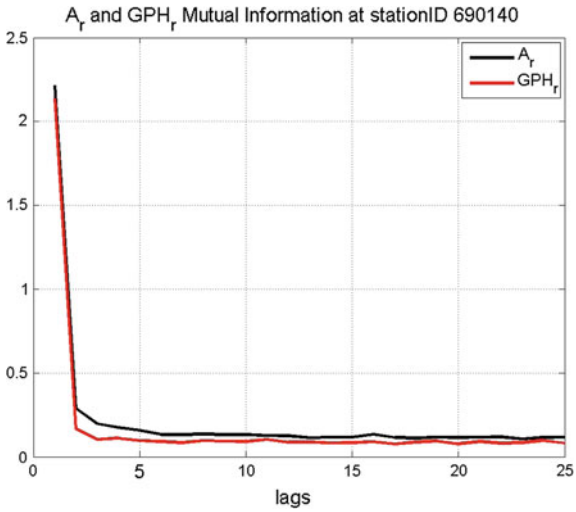

Fig. 7.3 Mutual information of Ar and GPH_r index computed the station ID726590

7.2 Clustering Daily Patterns of Solar Radiation

In order to evaluate the consistency of clustering the features described in the previous section, the silhouette (see Sect. 1.14.6.2) versus the cluster number was evaluated as shown in Fig. 7.4.

As it is possible to see the highest values of the silhouette are obtained by clustering into 2 classes. Nevertheless, it is possible to see that the silhouette is still significant also for classification into 3 and 4 classes.

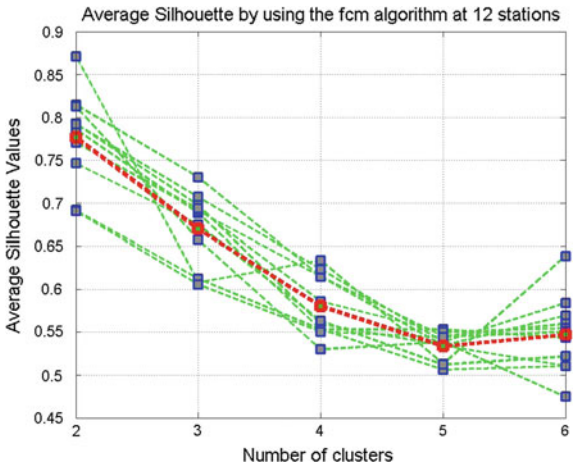

Fig. 7.4 Silhouette versus the number of clusters at 12 stations of the NOAA dataset; the *red curve* indicates the silhouette averaged over 12 stations

Clustering features into 3 and 4 classes using the *fcm* algorithm is shown in Fig. 7.5. As it is possible to see the centers are mainly distributed by increasing values of both A_r and GPH_r. Another aspect of clustering solar radiation daily patterns is that the cluster centers depend on the considered station, as shown in Fig. 7.6.

7.3 Daily Pattern Shapes

Solar radiation daily patterns belonging to the 4 classes, as classified by the *fcm* algorithm are show in Fig. 7.7. A heuristic description of these classes is the following:

- Class C_1: the class of completely cloudy sky days with big size clouds having a slow speed so that both the intensity of solar radiation and the dynamical level is weak, as shown in Fig. 7.7a.
- Class C_2: the class of days with significant sunshine combined with a large number of small clouds with high speed of passages and thus with high dynamic levels. An example is given in Fig. 7.7b.
- Class C_3: the class of days characterized by an important solar radiation with some clouds corresponding to a medium level dynamic as shown in Fig. 7.7c.
- Class C_1: the class of clear sky conditions days of solar radiation with very few clouds. An example is reported in Fig. 7.7d.

Since it is quite difficult to predict 1 day ahead the class of solar radiation using the features described above together with autoregressive models, we can try to draw some statistical information from the performed clustering.

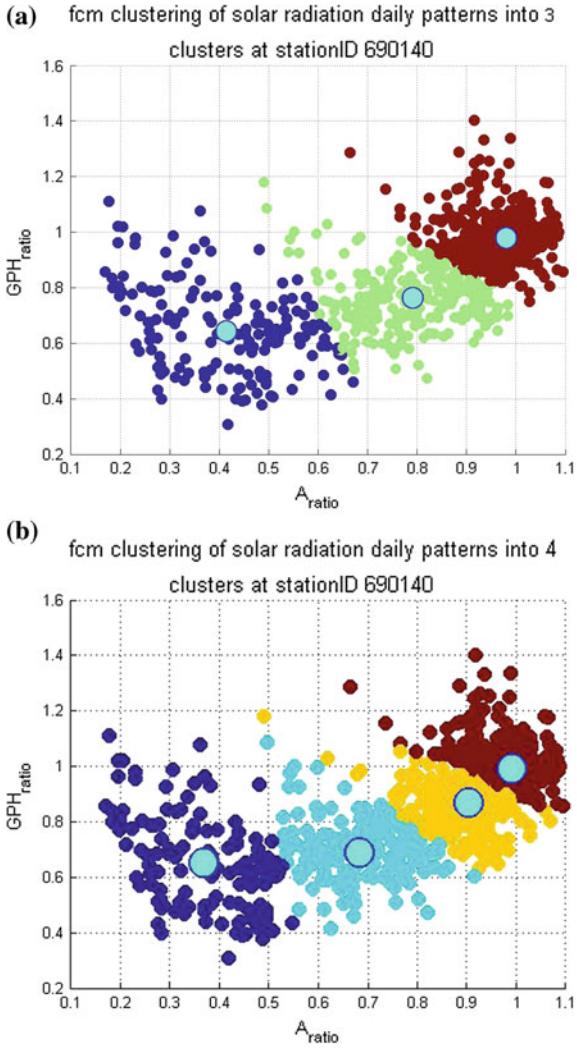

Fig. 7.5 Daily patterns and cluster centers computed for the station ID690140 **a** 3 clusters. **b** 4 clusters

7.3.1 Weight of a Solar Radiation Class

Here the weight W_i of a class C_i is defined as in (7.3)

$$W_i\% = \frac{n_i}{\sum_{i=1}^{c} n_i} 100 \tag{7.3}$$

where n_i is the number of patterns in class C_i and c is the number of considered classes. The knowledge of the class weight can provide some insights on how the daily patterns distributes during a determined time interval. Thus, for instance, in a

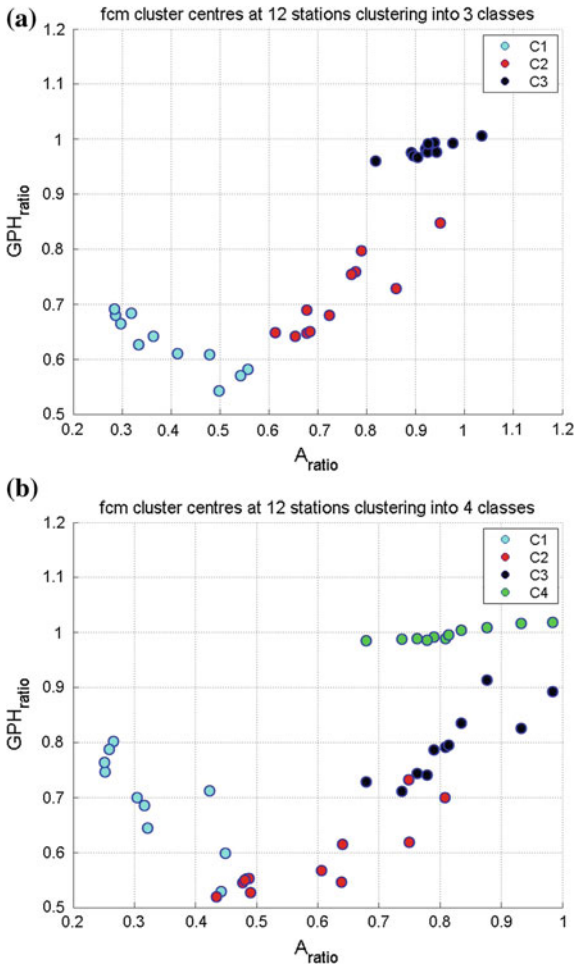

Fig. 7.6 Centers obtained clustering using the *fcm* algorithm at 12 stations **a** 3 clusters. **b** 4 clusters

4-class framework, the weight of each class, computed for the station ID 690140 using 3 years of data, as shown in Fig. 7.8. As it is possible to see the weight of the four classes are $C_1 = 11.5\,\%$, $C_2 = 14.2\,\%$, $C_3 = 23.4\,\%$, $C_4 = 50.9\,\%$, respectively. It is possible to see that at the considered station there is a clear prevalence of patterns in class C_4 (about 50 %). Since C_4 is the class characterized by completely clear sky conditions, we obtain an estimate of the convenience of using this form of renewable energy. Similarly, it is possible to observe that the weight of the class C_1, i.e., the class of completely cloudy solar radiation daily patterns, is that characterized by the lowest value (less than 12 %).

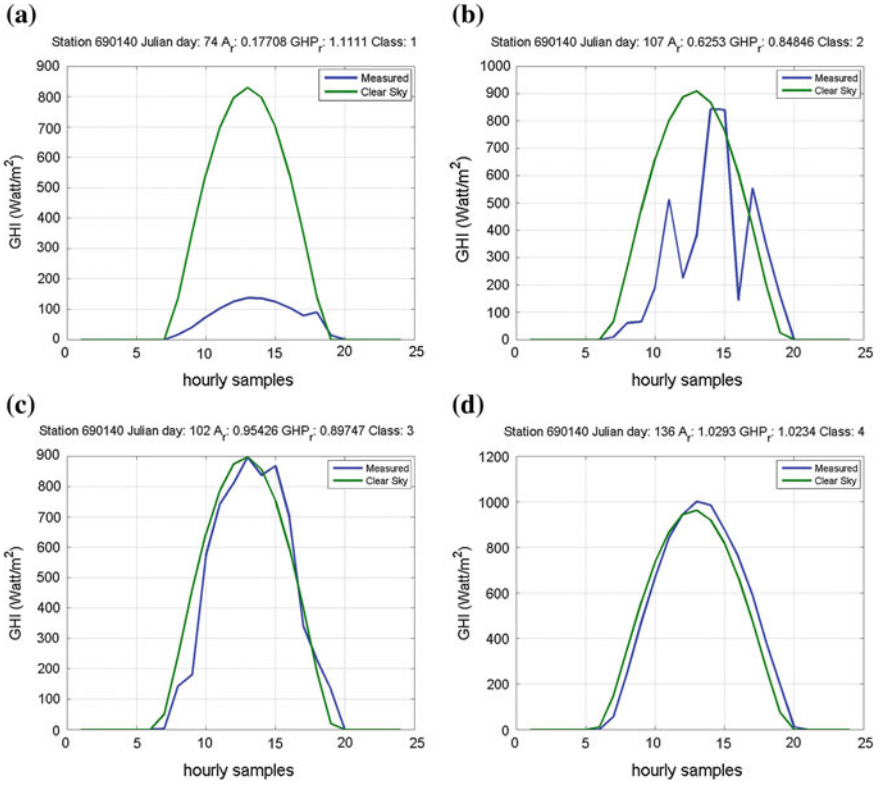

Fig. 7.7 Typical solar radiation daily patterns clustering into 4 classes **a** Class C_1. **b** Class C_2. **c** Class C_3. **d** Class C_4

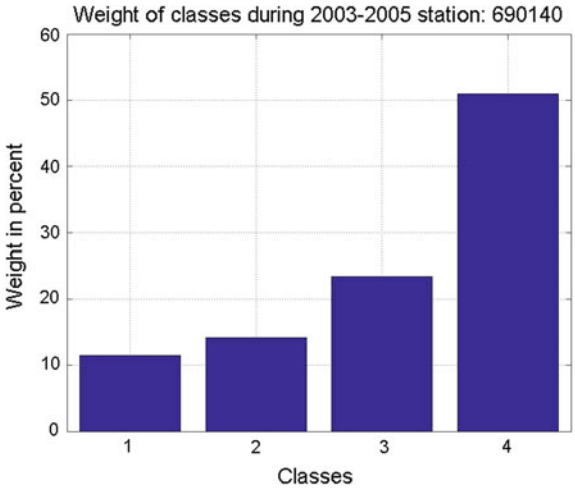

Fig. 7.8 Weight of solar radiation daily patterns, in a 4-class framework, computed at the station ID690140 using data recorded from 2003 to 2005

7.3.2 *Permanence of a Solar Radiation Class*

Clustering into four classes the overall solar radiation daily patterns recoded from 2003 to 2005 at the station ID690140, and computing the permanence, we get results shown in Fig. 7.9. It can be observed that the C_1 and C_2 classes, i.e., those characterized by low levels of daily solar radiation, in addition to being of low weight, as discussed above, exhibit a low degree of persistence within the respective classes. Instead the C_3 and especially the C_4 class, which are characterized by values of solar radiation similar to those recorded in clear sky conditions, exhibit a higher degree of persistence. In order to objectively characterize the degree of persistence $P_i(p)$ in

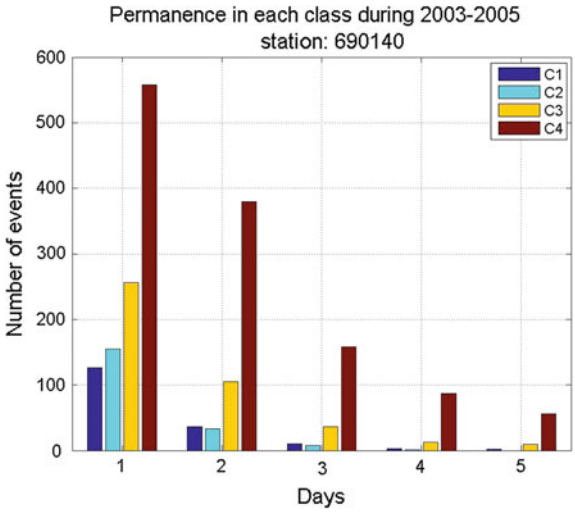

Fig. 7.9 Permanence of solar radiation daily patterns, in a 4-class framework, computed at the station ID 690140 using data recorded from 2003 to 2005

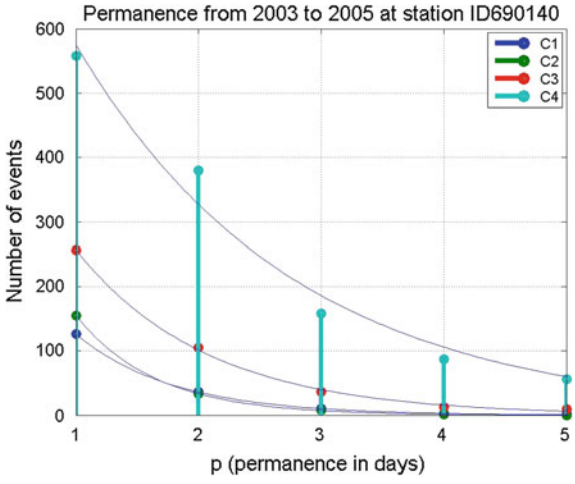

Fig. 7.10 Permanence and corresponding fit at station ID690140 during 2003–2005

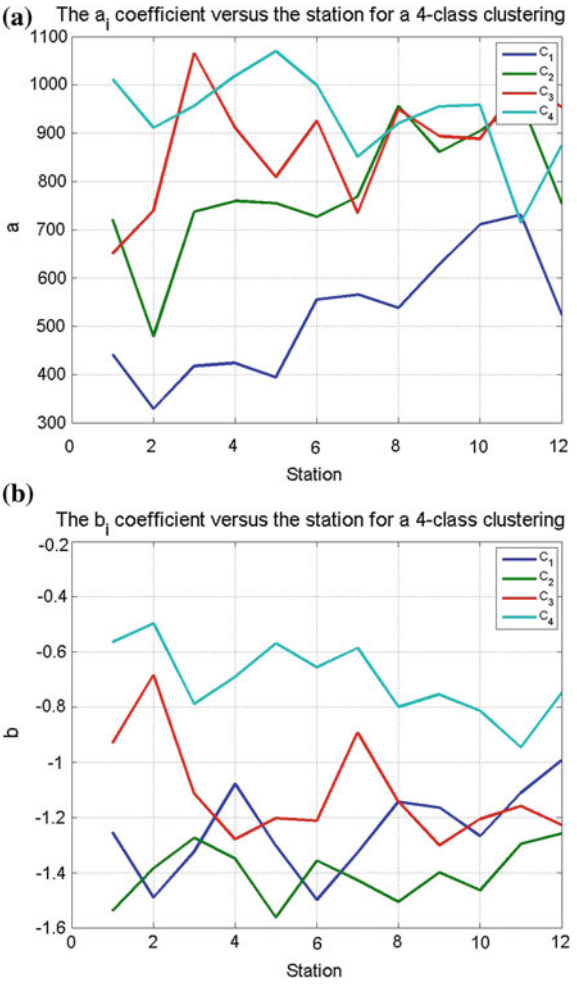

Fig. 7.11 Dependence of a and b on the class and station in a 4-class framework
a a-coefficient.
b b-coefficient

each class, we have tried to fit $P_i(p)$ by a simple decaying exponential of the form (7.4)

$$\begin{aligned} P_i(p) &= a_i \cdot \exp\left(b_i \cdot p\right), \\ p &= 1, 2, 3, 4, ..., \\ i &= 1, 2, ...c \end{aligned} \tag{7.4}$$

being a_i and b_i two constants that usually depend on the considered i_{th} class. In particular, the b_i coefficient represents the permanence rate of patterns in the i_{th} class.

As an example, clustering into four classes, the overall daily patterns recoded at the station ID690140 from 2003 to 2005, and computing the permanence, it is possible to get results shown in Fig. 7.10. In figure, the filled circles of different colors, indicate,

reading in the ordinate scale, the number of patterns in each class that persist a number of days, at least equal to p. Further, the solid lines indicate the fitting of $P_i(p)$ by continuous decaying exponentials, which allows an estimation of the a_i and b_i coefficients that appears in expression (7.4). The dependence of a_i and b_i with respect to the class and the recording stations are shown in Fig. 7.11. It is interesting to consider, above all, the variability of b_i with the class. Indeed, Fig. 7.11b shows that the permanence rate of patterns in class C_4 is greater than that of the other classes, not only relatively to the station ID690140 described above, but also in all the others considered stations.

7.4 Conclusions

In this chapter, we have proposed a feature-based clustering approach of solar radiation daily patterns. In particular, we have shown that clustering the pattern into four classes using the pair of proposed features is possible to recognize days characterized by completely cloudy sky (class C_1), days with sunshine and high dynamic levels (class C_2), days characterized by an important solar radiation level with a medium dynamic (class C_3) and, finally, the class of clear sky conditions (class C_4). Furthermore, we have shown how to calculate the weight and the persistence rate within each class. Results demonstrate that, regardless of the considered station, the C_4 and C_3, in the order, are those that have a greater weight and a greater rate of permanence. Such a kinds of statistical information can provide useful insights in the absence of reliable methods for 1-day ahead class prediction.

References

1. P. Ineichen, R. Perez, A New airmass independent formulation for the Linke turbidity coefficient. Phys. A **73**, 151–157 (2002)
2. R. Perez, A New Operational Model for Satellite-Derived Irradiances- Description and Validation. Sol. Energy **73**, 207–317 (2002)
3. J. Geweke, S. Porter-Hudak, J. Time Ser. Anal. **4**, 221 (1983)

Chapter 8
Clustering Daily Wind Speed Time Series

Abstract This chapter deals with the problem of clustering daily wind speed time series based on two features referred to as W_r and H, representing a measure of the relative average wind speed and the Hurst exponent, respectively. It is shown that using these features daily, wind speed patterns can be clustered into 2 or 3 classes and some useful statistical properties, such as the class weight and the persistence of patterns in a class, can be estimated.

8.1 Introduction

As discussed in Chap. 6, autoregressive models allow to predict hourly average wind speed with a limited accuracy, only at very short time horizon (a few hours). For this reason, the availability of alternative modeling techniques, such as those that refer to data mining and machine learning, may play a significant role to extract from time series some statistical information at daily scale. Previous work concerning application of machine learning approaches to wind speed was presented by [1], who suggested to consider the Markov chains to classify wind speed time series. Decision trees based on $if\text{-}then$ rules, have been proposed by [2] with the aim of implementing very short-term (1-h ahead) wind speed prediction models. Data mining techniques and clustering approaches to classify wind speed data in different cities of Turkey have been adopted by [3].

In this chapter, following a scheme similar to that adopted in the previous Chap. 7, we propose to clustering, daily wind speed time series based on two features described in the next Sect. 8.2. The data set considered for this work was taken from the Western Wind Resource Dataset (see Appendix A.2 for details).

8.2 Two Features of Daily Wind Speed Time Series

According to [4], who suggests do not work directly with the raw data, we introduce two features of daily wind speed time series, which are useful to represent in a two dimensional space the original high dimensional daily wind speed daily patterns.

L. Fortuna et al., *Nonlinear Modeling of Solar Radiation and Wind Speed Time Series*, SpringerBriefs in Energy, DOI 10.1007/978-3-319-38764-2_8

Such features were chosen based on the idea that one should represent a normalized measure of the daily average wind speed while the other should represent the correlation properties. The two features, referred to as W_r and H respectively are formally defined as described below.

8.3 The W_r Index

As it is known, one of the main features of wind speed is its irregular fluctuating nature which occurs at any time scale [5]. Thus, for instance, $10m$ average samples fluctuate with respect to the corresponding hourly average, hourly averages with respect their daily average and, again, daily averages with respect their weekly or monthly averages. As an example, fluctuations of daily average wind speed with respect to the corresponding monthly averages are shown in Fig. 8.1 which suggests that the monthly average could be considered as a normalizing factor for daily average wind speed. Of course, other choices are possible. Thus the W_r index, is formally defined as in expression (8.1)

$$W_r(d) = \frac{\bar{W}(d)}{\bar{W}_m} \quad d = 1, \ldots, gg(m), \quad m = 1, \ldots, 12. \tag{8.1}$$

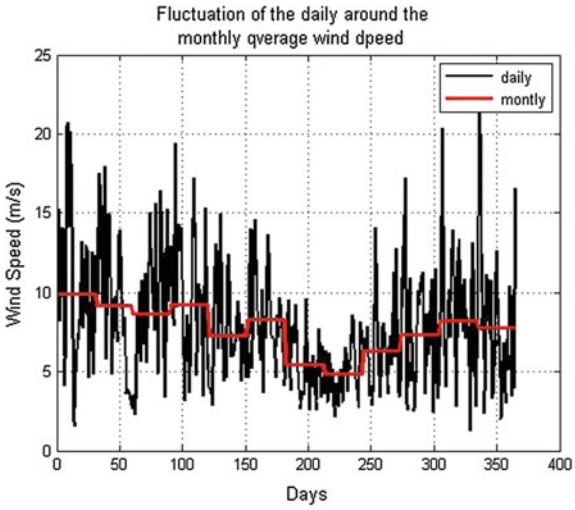

Fig. 8.1 Daily and monthly average wind speed

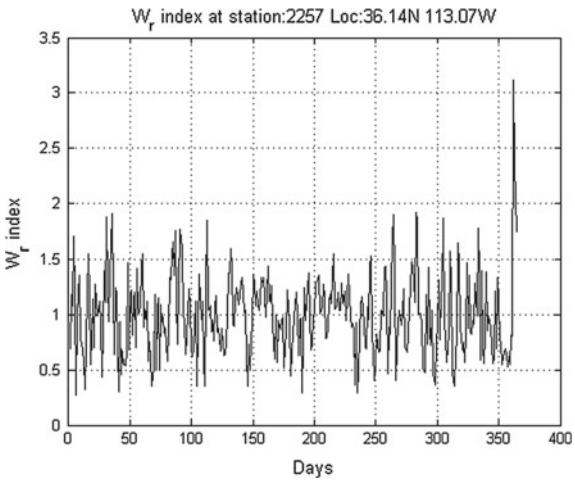

Fig. 8.2 W_r index computed during 1 year at the station ID2257

where:

- $\bar{W}(d)$ is the daily average wind speed at the generic day d;
- $\bar{W}_m$ is the monthly average wind speed at the month m to which d belongs; and
- $gg(m)$ is the number of days in the mth month.

Thus, the W_r index expresses a relative measure of the daily wind speed intensity. An example of W_r daily time series is shown in Fig. 8.2.

8.4 The Hurst Exponent of Daily Wind Speed

In order to represent the correlation properties of daily wind speed time series it was decided to consider the Hurst exponent, H. In particular, in this work the Hurst exponent was estimated using the technique, proposed by Geweke and Porter-Hudak (GPH) [6]. The Hurst exponent computed for the average 10-m daily wind speed time series recorded at the station ID2257, which means over time series of 144 values each, is shown in Fig. 8.3.

8.5 Clustering Wind Speed Daily Patterns

Based on the simulations carried out, we concluded that hierarchical classification approaches are not appropriate for the application we are dealing. The main reason, is that a time series of wind speed, even represented at daily scale by the two features, consists of hundreds or even thousands of pairs, i.e., patterns, to be classified. Indeed, since a multilevel hierarchy usually generates several clusters, that are not appropriate

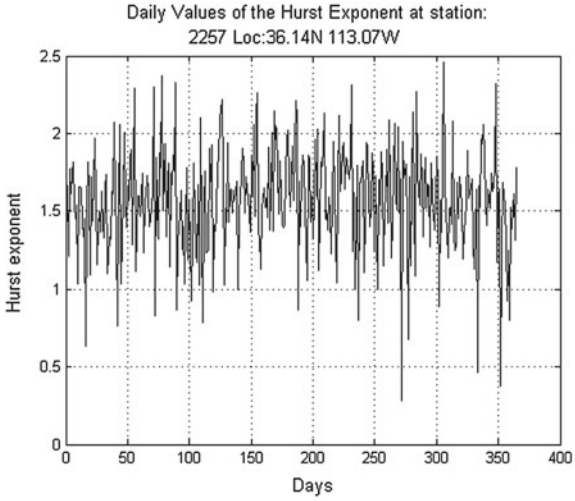

Fig. 8.3 H index computed during 1 year at the station ID2257

for wind speed daily patterns, in this work, we prefer to use a single-level hierarchy, such as that obtained using the *kmeans* of the *fcm* approaches. In particular, we choose the *fcm* algorithm as considered more flexible and stable with respect to the *kmeans*.

The classification approach described in this section consists of the two following steps:

1. Wind speed daily patterns, are mapped into pairs $(W_r(d), H(d))$.
2. The pairs $(W_r(d), H(d))$ are clustered by using the (*fcm*) algorithm.

Clustering examples of $(W_r(d), H(d))$ pairs into 2 and 3 classes using the *fcm* algorithm are shown in Fig. 8.4. As it is possible to see, clustering into 2 classes (see Fig. 8.4a), patterns essentially distributed following the W_r index, which thus plays the role of dominant feature. Roughly speaking it is possible to say that in a 2-class framework, class C_1 is represented by daily patterns featured by $W_r \leq 1$, while, of course, class C_2 by $W_r > 1$. Instead, for the 3-class framework (see Fig. 8.4b), a role in discriminating the daily wind speed patterns is also played by the H index.

Regarding the choice whether it is more appropriate to classify the daily patterns of wind speed in the two, three or more classes, of course it depends on the particular application and on the intrinsic features of patterns. From a strictly technical point of view, in order to evaluate the consistency of a particular choice, it is possible to refer to one of the criterion mentioned in Sect. 1.14.6. In particular, we have considered the silhouette criterion. As an example, the silhouette obtained clustering the features of station ID 2257 in 2 classes using the *fcm* algorithm, is shown in Fig. 8.5, which demonstrates that, for the considered station, the pair of wind speed features described in the previous Sect. 8.2, are generally well separated at least into two classes.

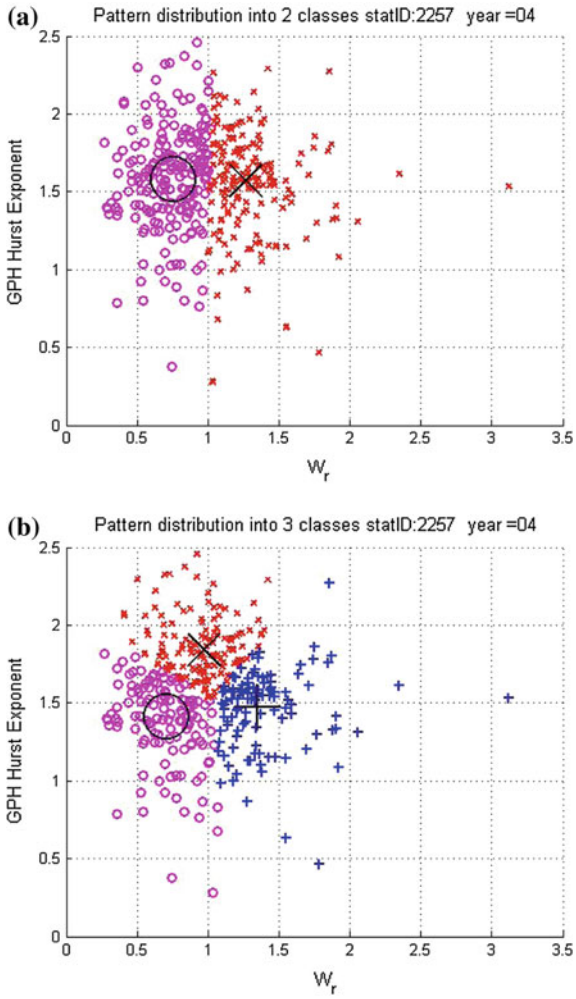

Fig. 8.4 Clustering the features of wind speed daily patterns into 2 and 3 classes, respectively, performed using the *fcm* algorithm. **a** 2 classes. **b** 3 classes

Nevertheless, in order to make choices that are as much as possible independent on the particular station, we have computed the average silhouette versus the number of clusters, averaging over twelve recording stations; results obtained are shown in Fig. 8.6. As it is possible to see, in the considered feature-based clustering problem the highest average value of the silhouette is obtained for the lowest number of classes, i.e., 2. Of course, this does not mean that it is not recommended to cluster in more than 2 classes, but simply that with a number of clusters greater than 2 the silhouette is on average lower than the maximum allowed.

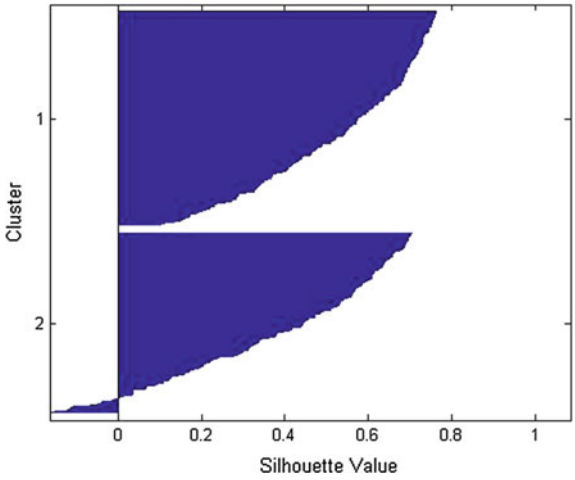

Fig. 8.5 Silhouette obtained clustering the features of the station ID 2257 in 2 classes by using the *fcm* algorithm

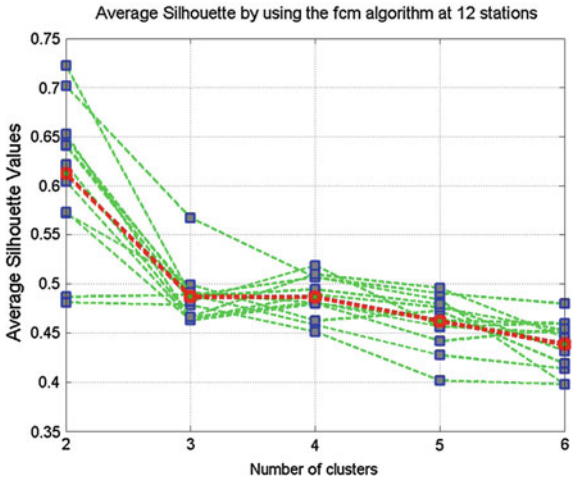

Fig. 8.6 Average silhouette versus the number of clusters at 12 recording stations (in *red* color the average silhouette)

8.5.1 Stability of the Wind Speed Features Cluster Centers

One of the questions that naturally arise in the described clustering problem is whether the cluster centers remain approximately constant from year to year and from station to station. We have experimentally found that, at least for 3-year time intervals, as is the case of the data set considered in this work, while the answer to the former question depends on the particular station, the answer to the latter question is certainly negative. To better illustrate the answer, we show in Fig. 8.7, the cluster centers computed yearly from 2004 to 2006 at the three stations, referred to as ID2257, ID6435 and ID 25766. As it is possible to see, the cluster centers are nearly coincident at the station ID 6435, they are partially coincident for the station ID 2257 and are

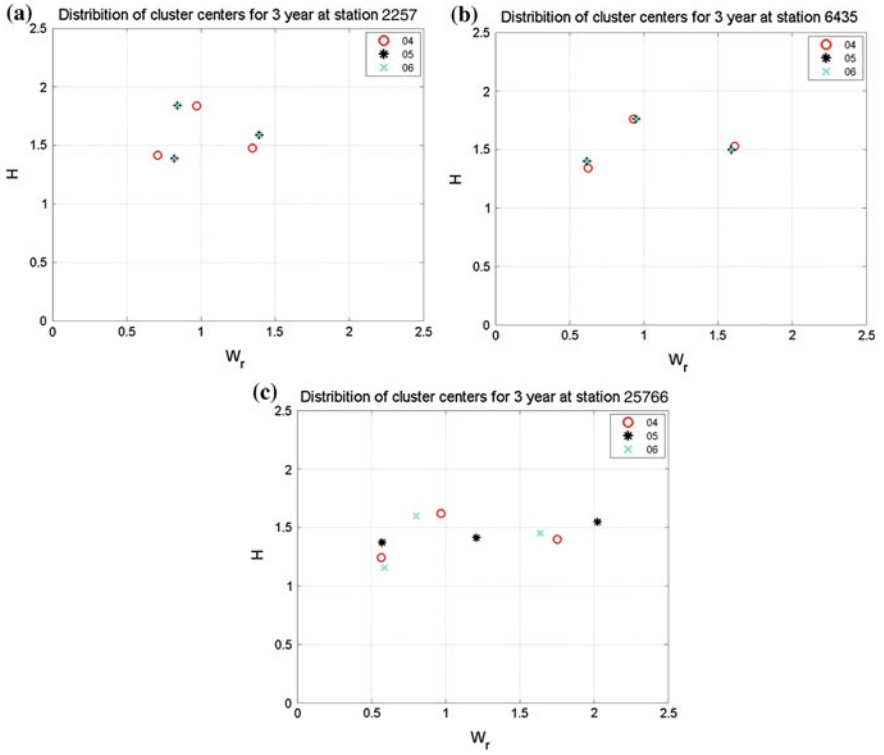

Fig. 8.7 Cluster centers computed yearly from 2004 to 2006 at three different stations. **a** station ID2257. **b** station ID 6435. **c** station ID 25766

not coincident at all for the station ID 25766. Of course, this finding could not be attributed solely to the physical characteristics of the wind speed time series, but it may also depend on the sensitivity to the initial conditions of the considered clustering algorithm. However, we have experimentally found that for most of the considered stations, the cluster centers computed using the *fcm* algorithm are almost coincident at least for two of the three considered years.

8.6 Some Applications

Once classes have been attributed to daily wind speed time series, some useful statistics can be carried out, such as computing the weight of each class during a predefined time interval or the permanence in days in each class, as explained in the following sections.

8.6.1 Weight of a Class

The weight W_i of a class C_i, is defined as in (7.3) of the previous Chapter devoted to clustering of solar radiation daily patterns, can provide some insights on how the daily average speed distributes during a predefined time interval. Thus, for instance, since at the station ID2257, during 2004, were classified $n_1 = 131$ patterns in class C_1, $n_2 = 166$ in class C_2 and $n_3 = 69$ patterns in class C_3, the class weights are $W_1\,\% = 35.80$, $W_2\,\% = 45.35$, and $W_2\,\% = 18.85$, respectively.

8.6.2 Permanence of Patterns in a Class

The permanence $P_i(p)$ in a class, already introduced in (7), expresses the number of patterns that, in predefined time interval expressed in days, remains consecutively in the same ith class. For instance, Fig. 8.8 shows that in a 3-class framework, at the station ID2257, during 2004, about 50 patterns persist in class C_3 at least 2 days, but less than 20 persist at least 3 days. We have experimentally found that it is possible to fit the permanence $P_i(p)$ of daily patterns by a simple decaying exponential of the form (8.2)

$$\begin{aligned} P_i(p) &= a_i \cdot \exp(b_i \cdot p), \\ p &= 1, 2, 3, 4, \ldots, \\ i &= 1, 2, \ldots c. \end{aligned} \tag{8.2}$$

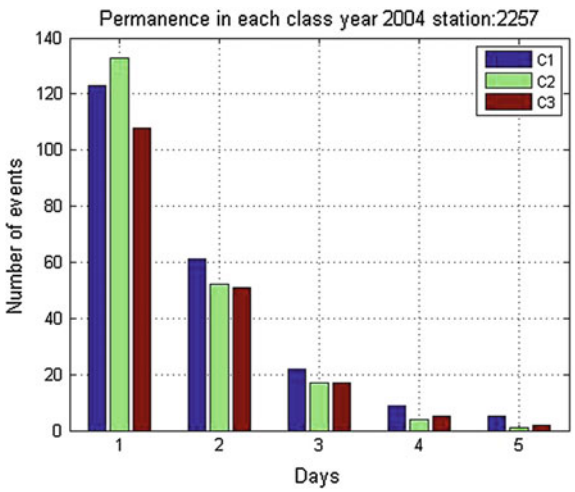

Fig. 8.8 Permanence (in days) in the same class at the station ID2257 in a 3-class framework

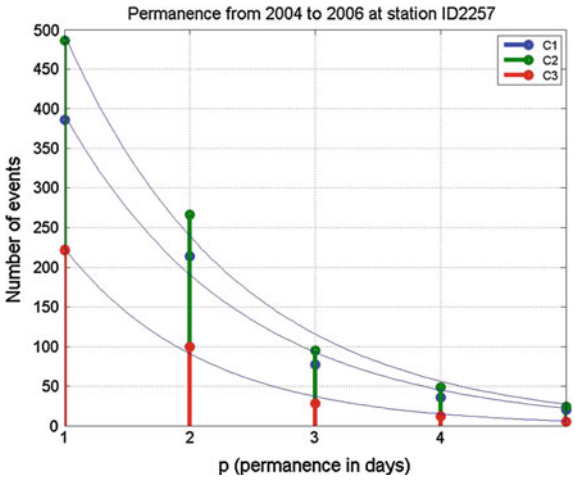

Fig. 8.9 Permanence and corresponding fitting at the station ID 2257 during 2004

being a_i and b_i two constants that usually depend on the considered ith class. In particular, the b_i coefficient represents the rate of permanence decaying of patterns in the ith class.

As an example, clustering into $c = 3$ classes the overall daily patterns recoded from 2004 to 2006 at the station ID2257, and computing the permanence, it is possible to get results as shown in Fig. 8.9. In the figure, the filled circles of different colors, indicate, reading in the ordinate scale, the number of patterns in each class that persist a number of days equal to p. Further, the solid lines indicate the fitting of $P_i(p)$ by continuous decaying exponentials, which allows an estimation of the a_i and b_i coefficients that appears in expression (8.2). Figure 8.10 shows that these coefficients depend on the recording station. It is interesting to consider, above all, the variability of b_i with the class. Indeed, Fig. 8.10b shows that the permanence rate of patterns in class C_1 is greater than that of the class C_2 which in turn is greater than that of the class C_3. This result can be related with the fact that the patterns of the classes C_1, C_2, and C_3 are sorted by W_r. So, in simple terms, the permanence of patterns characterized by relative high daily wind speed decay faster than patterns characterized by medium or low wind speed.

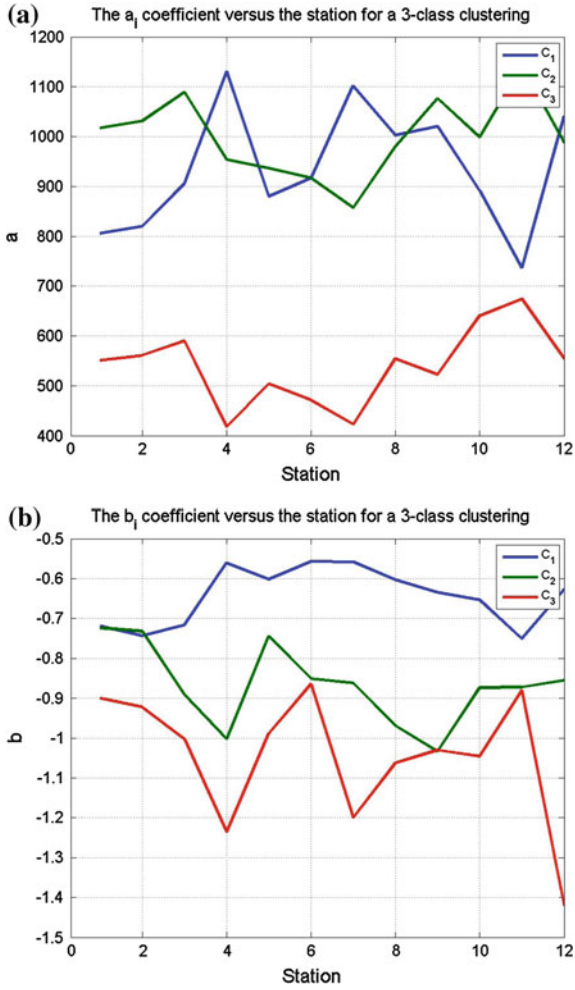

Fig. 8.10 Dependence of a and b on the class and station. **a** a-coefficient. **b** b-coefficient

8.7 Conclusions

In this chapter, a feature-based strategy to classify daily wind speed time series, based on a pair of indices referred to as W_r and H, has been proposed. Results demonstrate that the classification of wind speed daily patterns exhibit features that depend on the considered station, therefore confirming that the formation of wind speed process depends on the spatial coordinates where the data are collected. Although 1-day ahead prediction of the wind speed class is difficult to tackle using past wind speed data only, the clustering approach may help to gather useful information and contribute to the understanding of the studied phenomena.

References

1. Z. Song, X. Geng, A. Kusiak, C. Xu, Mining markov chain transition matrix from wind speed time series data. Expert Syst. Appl. **38**, 10229–10239 (2011)
2. A. Troncoso, S. Salcedo-Sanz, C. Casanova-Mateo, J. R., L. Prieto, Local models-based regression trees for very short-term wind speed prediction. Renew. Energy **81**, 589–598 (2015)
3. S.I. Colak, M. Demirtas, M. Yesilbudak, A data mining approach: analyzing wind speed and insolation period data in Turkey for installations of wind and solar power plants. Energy Convers. Manage. **65**, 185–197 (2013)
4. T.W. Liao, Clustering of time series data—a survey. Pattern Recogn. **38**, 1857–1874 (2005)
5. T. Laubrich, H. Kantz, Statistical analysis and stochastic modelling of boundary layer wind speed. Eur. Phys. J. Spec. Top. **174**, 197–206 (2009)
6. R. Weron, Estimating long range dependence finite sample properties and confidence intervals. Phys. A **312**, 285–299 (2002)

Chapter 9
Concluding Remarks

In this book, the problem of analyzing and modeling solar radiation and wind speed time series was addressed. This issues, as well as being a fascinating research topic, may have practical consequences in the problem of short term prediction, which is of great interest for managers of power plants. The studied prediction models are based on information that can be gathered from time series recorded at the site of interest only, thus excluding in the modeling process any other information, including the fact that the processes involved are spatial distributed.

The rationale for this choice is to obtain agile models, simple to design and implement, in contrast with the prediction models of type NWF (Numerical Weather Forecasting). On the other hand, a huge effort has been devoted in literature to implement prediction models based on time series only.

The first part of this book was devoted to perform a deep analysis of solar radiation and wind speed time series, not limited to the traditional stationary, spectral and autocorrelation analysis. Indeed, analysis carried out in Chaps. 2 and 3, was addressed to clarify several concerns of considered time series, including the hypothesis that the considered processes could be chaotic.

Despite the interpretation of chaos analysis results require further study, the others kinds of analysis performed, pointed out the complexity of physical phenomena underling solar radiation and wind speed time series and in particular that belong to the ubiquity class of $1/f$ noises or random walks, are long range correlated and exhibit multi-fractal spectrum.

Even if a better understanding of the issues relating to the presumably chaotic nature of these kinds of time series may led to more insights about physical aspects, nevertheless, the modeling technique adopted (i.e. the embedding phase-space representation) is the one most widely used for modeling of complex systems and therefore considered appropriate for the purposes of this book.

Results described show that while the proposed EPS based approach, represent a significant improvement, with respect to others popular reference models, for short term prediction of solar radiation, the advantages are much more limited for wind speed. This could be interpreted as a symptom of a greater degree of complexity of the phenomena related to wind speed with respect to solar radiation.

L. Fortuna et al., *Nonlinear Modeling of Solar Radiation and Wind Speed Time Series*, SpringerBriefs in Energy, DOI 10.1007/978-3-319-38764-2_9

Another results of this book is the attempt of using the cluster analysis in order to extract from hourly average time series some statistical information at daily scale. This could try to partially overcome the difficulties to predict 1 day ahead the averages of solar radiation and wind speed time series, due to the very serious limitations imposed by the autocorrelation of daily samples.

Clustering of daily patterns performed by using feature bases approaches, show that cluster features depends on the considered station. Nevertheless, classes can be clearly recognized and computed by using relative simple algorithms such as the *fcm*. In the book it was shown how useful information can be extracted, such as the weight of each class and the permanence of patterns in a given class. Nevertheless, these are just a few example of more deeply insights that can be performed by applying machine learning approaches to daily patterns of solar radiation and wind speed.

A limitation of the work carried out in this book was undoubtedly that the modeling was performed by considering time series recorded at individual recording stations, while the physical processes underling wind speed and solar radiation are distributed. This limitation is often not overcoming for the lack of recording stations in a given area. However, nowadays, since large distributed plants are available, it is realistic the possibility to involve in the modeling process correlation among time series recorded at different stations. This could improve the performances of statistical models described in this book.

Appendix A
Software Tools and Data

In this appendix, we provide the list of functions suggested trough the book and others useful web resources.

A.1 List of Functions

This is the list of software functions considered in this book. In parenthesis, it is indicated the corresponding package. Description of the packages is given in Sect. A.2.

adftest Perform the augmented Dicky–Fuller test (MATLAB® Econometrics Toolbox).
autocorr Perform the autocorrelation of a time series (MATLAB® Econometrics Toolbox).
boxcount Compute the fractal box dimension D (MATLAB File Exchange).
ClassificationTree Built a decision tree with binary splits for classification (MATLAB® Statistics Toolbox).
clusterdata Perform agglomerative clusters from data (MATLAB® Statistics Toolbox).
configure Configure network inputs and outputs to best match input and target data (MALAB® Neural Network Toolbox).
false_nearest Compute the fraction of false nearest neighbors (TISEAN project).
fcm Perform the Fuzzy C-means clustering (MATLAB® Fuzzy Logic Toolbox).
feedforwardnet Create a feedforward networks (MATLAB® Neural Network Toolbox).
fitdist create a probability distribution object (MATLAB® Statistics Toolbox).
detrend Remove a linear trend (MATLAB®).
evalclusters A class of function to evaluate clustering solutions (MALAB® Statistics Toolbox).
evalfis Perform fuzzy inference calculations (MATLAB® Fuzzy Toolbox).

L. Fortuna et al., *Nonlinear Modeling of Solar Radiation and Wind Speed Time Series*, SpringerBriefs in Energy, DOI 10.1007/978-3-319-38764-2

dfa Compute the Hurst exponent by using the DFA analysis (MATLAB File Exchange).
genfis3 Generate fuzzy inference system structure from data using fcm clustering (MATLAB® Fuzzy Logic Toolbox).
gmdistribution A class of functions which implement the Gaussian Mixture Clustering (MATLAB® Statistics Toolbox)
gph Compute the hurst exponent by using the Geweke–Porter–Hudak approach (MATLAB® File Exchange).
hurst Compute the hurst exponent by using the R/S analysis (MATLAB® File Exchange).
kmeans Perform the k-means clustering (MATLAB® Statistics Toolbox).
kpsstest Perform the Kwiatkowski–Phillips–Schmidt–Shin test (MATLAB® Econometrics Toolbox).
lazy Perform noise reduction by locally constant approximations (TISEAN project).
lyap_k Compute the maximal Lyapunov exponent (TISEAN project).
mutual Compute the mutual information (TISEAN project).
MFDFA Perform the Multi-fractal detrended fluctuation analysis (MATLAB File Exchange).
pdf Return probability density function (MATLAB® Statistics Toolbox).
periodogram Estimate power spectral density (MATLAB® Signal Processing Toolbox).
plotsomhits Plot a SOM hit matrix (MATLAB® Neural Network Toolbox).
pptest Phillips–Perron test for one unit root (MATLAB® Econometrics Toolbox).
pvl_clearsky_ineichen Compute the global horizontal irradiance in clear sky condition bases on the Ineichen and Perez model (PVLib Toolbox).
recurr Compute the Recurrent plot (TISEAN project).
train Train neural network (MATLAB® Neural Network Toolbox).
selforgmap Configure a Self-Organized Map (MATLAB® Neural Network Toolbox).
silhouette Perform the Silhouette plot (MATLAB® Statistics Toolbox).
spectrum Perform the power spectral density estimate (TISEAN project).
vratiotest Variance ratio test for random walk (MATLAB® Econometrics Toolbox).

A.2 Tools and Data

This is the list of software tools and data considered trough the book.

MATLAB is a commercial tool for numerical computation and visualization. It consists of several Toolboxes. The interesting ones, for the purpose of this book are the Neural Network, the Fuzzy, the Statistic, the Signal Processing and the Econometric Toolboxes.
MATLAB Central is a site which allows to exchange MATLAB code developed by various authors. Files can be free download. The code, together with terms and

conditions for its use, can be found in http://it.mathworks.com/matlabcentral/about/fx/.

TISEAN is a free software project for the analysis of time series with methods based on the theory of nonlinear deterministic dynamical systems, or chaos theory. Software programs developed in the framework of this project are described in [1–3]. The software package, together with terms and conditions for its use, is available for downloading from http://www.mpipks-dresden.mpg.de/~tisean/.

PVLib Toolbox The clear sky model considered in Chap. 5 which implement the Ineichen and Perez model for global horizontal irradiance as presented in [4, 5] is part of the MATLAB_PVLib Toolbox, download from the Sandia National Labs PV Modeling Collaborative (PVMC) platform.

NREL The data set of hourly average time series considered in this book are part of the National Solar Radiation Database, managed by the NREL (National Renewable Energy Laboratory) of USA. Data of this database was recorded from 1999 to 2005 and can be freely downloaded from ftp://ftp.ncdc.noaa.gov/pub/data/nsrdb-solar/.

WWR The data set of wind speed time series considered in this book is a subset of the Western Wind Resource (WWR) dataset, a large database which stores data of more than 30, 000 sites that were modeled in the framework of Western Wind and Solar Integration Study. Data are at the present public available from http://wind.nrel.gov/Web_nrel/.

References

1. R. Hegger, H. Kantz, T. Schreiber, Practical implementation of nonlinear time series methods: the TISEAN package. Chaos **9**, 413 (1999)
2. H. Kantz, T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, 1997)
3. T. Schreiber, A. Schmitz, Surrogate time series, Phys. D, 142–346 (2000)
4. P. Ineichen, R. Perez, A New airmass independent formulation for the Linke turbidity coefficient. Phys. A **73**, 151–157 (2002)
5. R. Perez, A new operational model for satellite-derived irradiances-description and validation. Solar Energy **73**, 207–317 (2002)

Index

L. Fortuna et al., *Nonlinear Modeling of Solar Radiation and Wind Speed Time Series*, SpringerBriefs in Energy, DOI 10.1007/978-3-319-38764-2