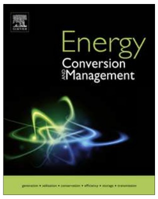
Energy Conversion and Management 75 (2013) 561–569

Contents lists available at SciVerse ScienceDirect

Energy Conversion and Management

journal homepage: www.elsevier.com/locate/enconman



Forecasting hourly global solar radiation using hybrid *k*-means and nonlinear autoregressive neural network models

Khalil Benmouiza [a,1], Ali Cheknane [b,*]

[a] Département de Physique, Faculté des Sciences, Université Abou-Beker Belkaid de Tlemcen, BP 119, Tlemcen 13000, Algerie
[b] Laboratoire des Semiconducteurs et Matériaux Fonctionnels, Université Amar Telidji de Laghouat, Algérie, BP 37G, Laghouat 03000, Algerie

ARTICLE INFO

Article history:
Received 14 May 2013
Accepted 5 July 2013

Keywords:
Forecasting solar radiation
Artificial neural networks
Clustering
Phase space reconstitution

ABSTRACT

In this paper, we review our work for forecasting hourly global horizontal solar radiation based on the combination of unsupervised *k*-means clustering algorithm and artificial neural networks (ANN). *k*-Means algorithm focused on extracting useful information from the data with the aim of modeling the time series behavior and find patterns of the input space by clustering the data. On the other hand, nonlinear autoregressive (NAR) neural networks are powerful computational models for modeling and forecasting nonlinear time series. Taking the advantage of both methods, a new method was proposed combining *k*-means algorithm and NAR network to provide better forecasting results.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

The generation of the energy in our modern industrialized society is still mainly based on a very limited resource. Some projections show that the global energy demands will almost triple by 2050 [10]. Thus, the search for alternative energy resources has become an important issue for our time. Solar energy is becoming a very attractive solution since it is considered an essentially inexhaustible and broadly available energy.

For an efficient conversion and utilization of solar power, solar radiation data should be measured continuously and accurately over the long-term. However, the measurement of solar radiation is not available for all countries in the world due to some technical and fiscal limitations. Hence, several studies were proposed in the literature to find mathematical and physical models to estimate and forecast the amount of solar radiations such as stochastic prediction models based on time series methods [14,36,39,40] and artificial neural network approaches [11,2,4].

Classical linear time series models like autoregressive moving average modeling [3] have been widely used in modeling of linear time series [36]. Even so, it was proven that they are inadequate in the analysis and prediction of solar radiation due to the non-stationary and nonlinearity of the solar radiation time series, especially for cloudy sky [39,1,36]. In addition, stochastic models are based on the probability estimation that needs a full identification of the mathematical function, leads to a difficult forecasting of the solar radiation time series [36]. Moreover, global solar radiation time series is a dynamical system that depends on some meteorological elements such as temperature, water vapor, suspend solids, cloud and water air condition that can represent nonlinear characteristics [36,12,38] .

To overcome this problem, nonlinear approaches, such as artificial neural networks (ANN) was considered a powerful tool for forecasting similar time series [39,28,36]. The advantages of the ANN that it does not require the knowledge of the internal system parameters that offer a compact solution for multiple variable problems [11,2,36,4]. However, single models presented a big forecasting error [36]. Thus, hybrid methods combining different models have been widely used in the literature to improve the forecast performance [36,12,5]. Nevertheless, no one of those methods will be capable of presenting information about the behavior of the solar radiation time series in the future. Hence, it was used the Time Series Data Mining (TSDM) methodology [33] which is a fundamental contribution to the fields of time series analysis and data mining that allows a search, for valuable information on nonlinear problems such as solar radiation time series [21].

Data mining is the identification of interesting structure in the data, where the structure designates patterns of the data and relationships among regions of the data; it is a process of grouping similar elements gathered closely using unsupervised clustering methods such as *k*-means and *c*-means algorithms [37]. Data mining techniques were used in a wide variety of fields for prediction. For example, in stock prices, meteorological data, customer behavior, production control and other types of scientific data [9].

Taking the two advantages of both methods, the *k*-means approach [25] for clustering the solar radiation data to extract useful

[*] Corresponding author. Tel.: +213 (0)29 93 21 17; fax: +213 (0)29 93 26 98.
E-mail addresses: benkhalil3@gmail.com (K. Benmouiza), cheknanali@yahoo.com, a.cheknane@mail.lagh-univ.dz (A. Cheknane).

[1] Tel.: +213 (0)778 73 85 56.

0196-8904/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.enconman.2013.07.003

information and the ANN for forecasting purposes, a new method was proposed in this paper that combines an unsupervised k-means clustering algorithm and nonlinear autoregressive neural network.

At the first stage, the data obtained from the phase space reconstitution using Takens theorem [35] were clustered using k-mean algorithm; clustering is a process of grouping an unlabelled set of examples into several clusters such that a similar pattern is associated with every cluster. The motivation of using the k-means approach in this paper is due to its simplicity and also to the fact that the proposed methods do not require an advanced clustering algorithm. However, one of the vital issues of the k-means algorithm is the choosing of the appropriate number of clusters [37]. Therefore, a silhouette function proposed by [32,23] was used to obtain the best number of bunches.

At the second stage, the nonlinear autoregressive (NAR) neural network that is a multilayer perceptron neural network (MLP) with some modification was applied for forecasting the solar radiation time series trying different architecture to get the best network structure. Combining those two methods presented better results for multi-step ahead prediction in long term forecasting.

The remaining part of this paper is organized as follows. Section 2 presented the methodology used in this work for forecasting the solar radiation time series using time series data mining technique, a background of space phase reconstitution, k-means clustering algorithm and NAR network methods were also viewed. In Section 3, we simulated the forecasting results of the proposed method and comparing the results with the measured ones. The last section was devoted to the conclusion and discussion of future works.

2. Methodology

A time series is a collection of time ordered observations $x(t_i)$, each one being entered at a specific time t called a period [30]. Modeling and forecasting of the time series are an importation task to extract useful information from the data [7]. Hence, in this paper, a proposed method that relies on principles of time series analysis, unsupervised clustering, artificial neural networks and evolutionary optimization methods were proposed as presented in Fig. 1. The methodology can be outlined in the following steps:

(1) Determine the minimum, appropriate, embedding dimension for phase space reconstruction for the time series [15];
(2) Identify regions of the reconstructed phase-space which has similar characteristics using k-means clustering algorithm;
(3) For each cluster train different NAR neural network to generate regional predictor for forecasting local regions;
(4) Use the corresponding NAR neural network using different delay and neurons to generate a global prediction for the time series;
(5) Reconstructed phase-space of the obtained time series from step 4, then use the appropriate k-means method to cluster the data using the same parameters used in step 1 and step 2;
(6) To perform the forecast, assign each pattern from step 5 to the appropriate region obtained from step 3 using as a criterion the Euclidean distance:
- If the Euclidean distance between each region and the assigned pattern is small, then it was considered a better forecast, else return to step 4.

2.1. Determining an appropriate embedding dimension

Phase space reconstruction provides a simplified, multidimensional representation of a nonlinear time series that simplifies further analysis. The approach of phase-space reconstruction consists of embedding the time series into a higher-dimensional space to see the underlying dynamical system [15]. The most widely used version of embedding is a time delay embedding [35]. This method

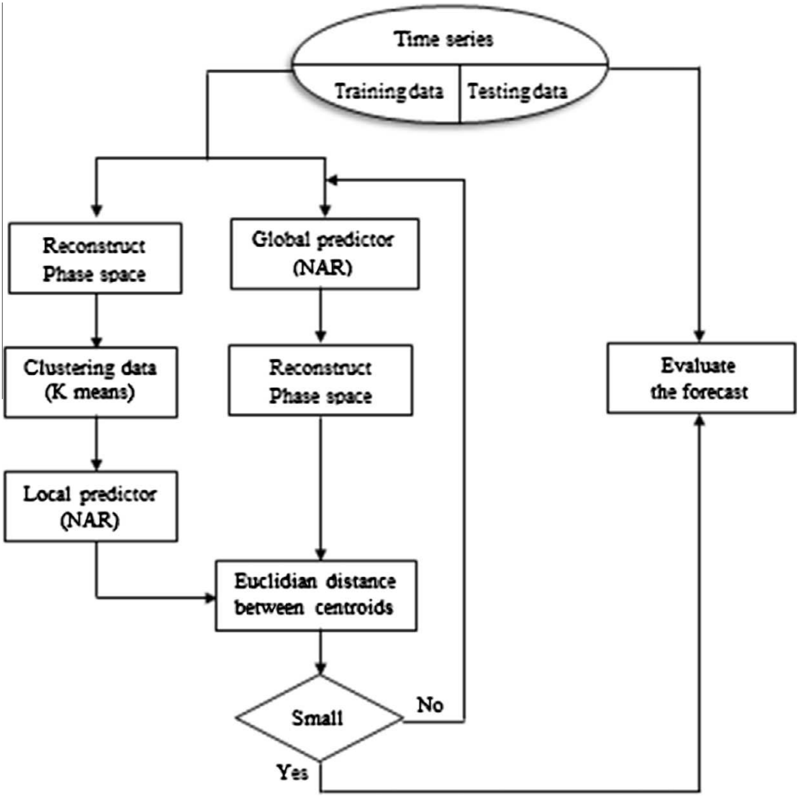


Fig. 1. The proposed methodology for time series data mining forecasting.

embeds a scalar time series $x(t_i)$ into a m-dimensional space denoted $X(t_i)$, as expressed in the following equation,

$$X(t_i) = (x(t_i), x(t_i + \tau), \cdots, x(t_i + (m-1)\tau) \tag{1}$$

where $i = (1, 2, \ldots, M)$, τ is the delay time, m is the embedding dimension, and M is the number of embedded points in the m-dimensional space given by Eq. (2). N is the total number of points of the time series and $X(t_i)$ is the embedded time series into an m-dimensional space.

$$M = N - (m-1)\tau \tag{2}$$

Several methods were presented in the literature to provide an estimation of optimal embedding dimension and time delay for better phase space reconstitution of the original time series [35,16,31]. In this paper, the mutual information method proposed by Fraser and Swinney [8] was used to set the delay coordinates. This method is summarized as follows,

– Calculating of the mutual information $I(x(t), x(t-\tau))$ of $x(t)$ and $x(t-\tau)$ for a given τ as expressed in the following equation,

$$I(x(t), x(t-\tau)) = \sum_{x\in\chi}\sum_{y\in\gamma} p(x(t), x(t-\tau)) \times \log \frac{p(x(t), x(t-\tau))}{p(x(t))p((t-\tau))} \tag{3}$$

$p(x(t), x(t-\tau))$, is the joint probability mass function for the marginal probability mass functions $x(t)$ and $x(t-\tau)$.

- Drawing of the mutual information function $I(t)$ for given τ,
- The optimum time delay τ is the first minimum of the mutual information function.

A small value of the delay leads to a $x(t)$ very similar to $x(t+\tau)$ then all the data stay near one other. On other hands, big delay leads to an independent coordinates and no information can be gained from the plotted data.

To determine the optimal embedding dimension m, different methods such as the box-counting dimension [26], false nearest neighbors [15], small-window solution [18] and C–C methods [16] were proposed in the literature.

In this paper, false nearest model was employed because of simple implementation and accuracy. It consists of learning how many dimensions are sufficient to embed a particular time series [15]; for a given embedding dimension, this method determines the nearest neighbor of every point in a given dimension, then checks to see if these are still close neighbors in one higher dimension. The percentage of False Nearest Neighbors should drop to 0 when the appropriate embedding dimension has been achieved.

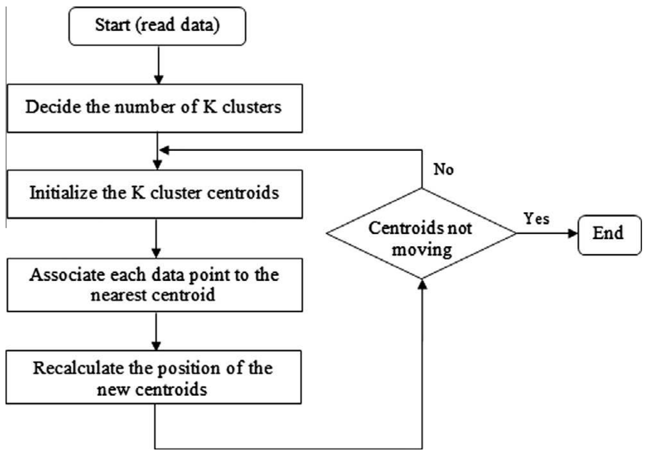

Start (read data)
Decide the number of K clusters
Initialize the K cluster centroids
Associate each data point to the nearest centroid
Recalculate the position of the new centroids
Centroids not moving
No
Yes
End

Fig. 2. *k*-Means clustering algorithm.

2.2. k-Means algorithm

k-Means is one of the quickest and simplest unsupervised learning algorithms to perform clustering; the method consists of classifying a given data into fixed *k* clusters [25,34]. The main idea is to define *k* centroids for each cluster; those centroids should be placed as much as possible far away from each other. In first step, each point of the data set is connected to the nearest cluster centroid by calculating the squared Euclidian distance between data point $x_i^{(j)}$ and the cluster centre c_j, as expressed by the following equation

$$\|x_i^{(j)} - c_j\|^2 \tag{4}$$

The second step consists of re-calculating the location of the new *k* centroid. Repeating the first and second steps until the centroids no longer move produced a separation of the objects into groups from which the objective function *J* expressed in Eq. (5) is minimized.

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n}\|x_i^{(j)} - c_j\|^2 \tag{5}$$

A summary of *k*-means algorithm is shown in Fig. 2,

2.2.1. Selection of the number of clusters

The *k*-means algorithm is based on the selection of the optimum number of clusters [34,37]. The choosing of many clusters does not necessarily imply having a better quality of information. On the other hand, a small number of clusters produce unclear results that could muddle the pattern recognition up.

The Silhouette function [32] expressed in Eq. (6) provides a measure of the cluster separation that can be used for the interpretation and validation of clustered data. The motivation of using this technique that is simple to read, and provides a graphical representation that allows the testing of various sets of clusters. It consists of calculating the average dissimilarity $a\ (i)$ of the *i*th data within the same cluster. This criterion can be interpreted as how well-matched the *i*th data to those clusters are assigned to it. The next step, is to determine the average dissimilarity of the *i*th data with the data of another cluster, then the lowest average is denoted by $b(i)$.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{6}$$

From this equation, it is clearly shown that if $s(i)$ is close to 1 then $a(i) \ll b(i)$, which means that the values of $a(i)$ are too small, which indicate that the *i*th data is well matched for its cluster. Furthermore, a large $b(i)$ implies that i is badly matched to its neighboring cluster. Thus, a $s(i)$ close to 1 means that the datum is appropriately clustered. If $s(i)$ is close to minus one, then by the same logic, we can see that i would be more appropriate if it was clustered in its neighboring cluster. An $s(i)$ near zeromeans that the datum is on the border of two natural clusters.

A successful clustering has a high mean silhouette value $s(i)$. Lletí et al. [23] considered a 0.6 silhouette value for all clusters as a good result. However, in real-time series, it is almost impossible to achieve this. Hence, a compromise among silhouette plots and averages was used to determine the natural number of clusters within a data set.

2.3. Nonlinear autoregressive neural network (NAR)

Artificial neural network (ANN) is a class of neural network represented by a mathematical model that is inspired by the biological nervous system; it is an intelligent system that has the ability to recognize time series patterns and nonlinear characteristics.

Hence, it has been widely used for modeling dynamic nonlinear time series [13,22].

ANN combines artificial neurons to process information; it is made up by simple neurons that are connected in a network by weighted links. Each input is multiplied by those weights that computed by a mathematical function which defines the activation of the neuron. Another activation function computes the output of the artificial neuron that depends on a certain threshold.

Using mathematical notation, the output of a neuron can be written as the following equation,

$$y = f\left(b + \sum_i w_i x_i\right) \tag{7}$$

here b is the bias for the neuron; the bias input to the neuron algorithm is an offset value that helps the signal to exceed the activation function's threshold. f is the activation function, w_i are the weights, x_i are the inputs and y represents the output.

Various types of artificial neural networks were presented in literature among them Multi-Layer Perceptron (MLP), where the neurons are grouped into an input layer, one or more hidden layers and an output layer. Recurrent Neural Networks (RNN) such as layer recurrent networks [13], Time Delay Neural Networks (TDNN) [13,36] and NAR [6,27]. In RNN, the outputs of a dynamic system depend not only on the present inputs, but also on the history of the states systems and the inputs. The NAR is a recurrent dynamic network based on a linear autoregressive model with feedback connections, including several layers of the network. It is commonly used in multi-step ahead time series forecasting; it uses past values of the actual time series to predict next values as determined by the following equation,

$$\hat{y}(t) = f(y(t-1) + y(t-2) + \cdots + y(t-d)) \tag{8}$$

f is a nonlinear function, where the future values depend only on regressed d previous values of the output signal as shown in Fig. 3. The combined history of the inputs and outputs of the system forms an intermediate inputs vector to be shown in the neural network model that could be any of the standard feed forward neural networks like MLP networks.

In addition, the RNN are based on training algorithms that used to adjust the weight values to get a desired output when certain inputs are given. Hence, various ways were presented to let a neural network learn such as supervised training where the input–output set is defined, and unsupervised learning that the output is undefined.

Back-propagation method is one of the most popular and widely used learning techniques for training RNN. It consists of minimizing the global quadratic error between the network output and the desired target by adjusting the weight values. The adjustment can be done using several algorithms such as Levemberg–Marquardt [19,25], Bayesian Regularization [24] and scaled conjugate gradient [29] algorithms. The latter one was selected to train larger networks. Once the network is trained using the preselected inputs and outputs, all the synaptic weights are saved, and the network is ready to be tested on the new input information. Since the NAR network is very similar to a Multilayer Perceptron (MPL), a modified MLP neural network was applied in this paper for predicting purposes.

3. Simulation results

In our simulation, we are interested in multi-hour ahead forecasting of the hourly global solar radiation time series using a combination of clustering techniques and nonlinear autoregressive neural networks. Hence, two global horizontal solar radiation time series were selected in this paper for simulation purposes. In all cases, the evaluation of the accuracy of the prediction methodology is accomplished by calculating the root mean square error (RMSE) expressed by Eq. (9) and the normalized root mean square error (NRMSE) given by Eq. (10),

$$\mathrm{RMSE} = \left[< (I_{i,predicted} - I_{i,measured})^2 >\right]^{\frac{1}{2}} \tag{9}$$

$$\mathrm{NRMSE} = \left(\frac{\left[< (I_{i,predicted} - I_{i,measured})^2 >\right]^{\frac{1}{2}}}{< I_{i,measured} >}\right) \tag{10}$$

RMSE and NRMSE provide information on the short-term performance of the correlations by allowing a term-by-term comparison of the actual difference between the predicted and measured values. An NRMSE value between 0.2 and 0.5 was considered by Lewis [20] to be as a good prediction model. Kostylev and Pavlovski [17] found that the best performing model on an hourly time scale had an NRMSE of 0.17 for mostly clear days and 0.32 for mostly cloudy days. Furthermore, Wu and Chan [36] found that the NRMSE error will be big in the case of cloudy skies.

In addition, a comparison between the introduced naïve autoregressive and moving average (ARMA) predictors was used to evaluate the goodness of the proposed method. ARMA model has been widely used in papers for forecasting solar radiation time series [36]. It consists of modeling a time series of its past values, as expressed in the following equation,

$$x_t = \sum_{i=1}^{p} \varphi_i x_{t-i} + e_t + \sum_j^{q} \theta_j e_{t-j} \tag{11}$$

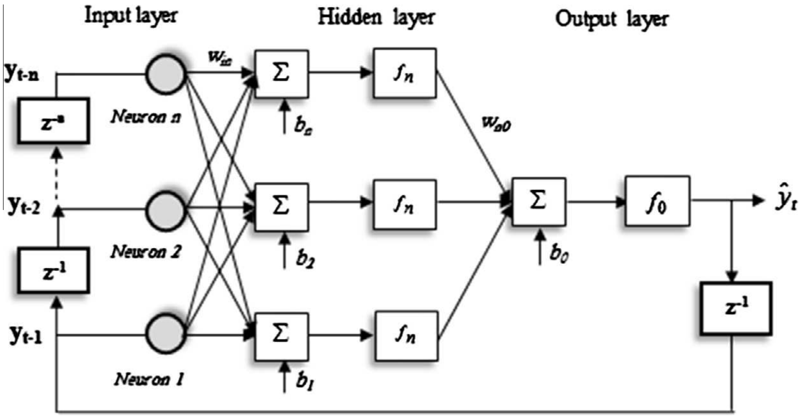


Fig. 3. Structure of NAR network.

where $\phi_i(i = 1\ldots p)$ and $\theta_j(j = 1\ldots q)$ are the constants representing the autoregressive AR, and the moving average MA parameters of order p, q, respectively. x_t is the actual value and e_t represents the Gaussian white noise with the mean zero in time t. The Box Jenkins methodology [3] was used to define the parameters of Eq. (11).

First, the average monthly global horizontal solar radiation time series between the years 1994 to 1996 was used. This time series takes the average of solar radiation for each hour during a month. The data were compiled from the National Meteorological Office of Algeria for the site of Oran, as shown in Fig. 4. The data set was divided into two samples of training data (from January 1994 to June 1996) and testing data (from July 1996 to December 1996), the training data set was used exclusively for model development then the test sample was used to evaluate the established model.

The first step in the analysis and prediction of this time series is the choice of an appropriate time delay, τ and the determination of the embedding dimension, m. To select τ an established approach is to use the value that yields the first minimum of the mutual information function; a time delay of 1 and an embedding dimension equal to 2 were used in the simulation.

The next step is to apply the k-means algorithm for clustering the high dimensional training data set obtained from phase space reconstitution in the previous step. At each step, different numbers of clusters were examined; the silhouette function was calculated for the determination of the right number of clusters. The metric used was squared Euclidean distance. We had plotted difference silhouette functions for the average monthly global horizontal solar radiation with a different number of clusters, and we established that the choosing of three clusters to be the best choice as shown in Fig. 5.

From Fig. 5, it is clearly shown that the most of the mean silhouette values are high. However, it is nearly impossible to arrive at a value of 0.6 for all clusters and not having negative values as the case of the first cluster. Thus, the appropriate number of clusters is usually taken when the graphical representation provides satisfactory results that mean when most of the silhouette values are high as expressed in Lletí et al. [23].

The obtained three cluster groups the solar radiation time series into three categories, high values of solar radiation which represent the noon hours, medium values that represent the day hours from 9 to 11 o'clock (or sky with medium clouds) and low solar radiation values, which represent hours of sunrise and sunset (or the presence of clouds).

After that, the NAR method with different architectures was applied to generate local predictor for each cluster that provides what we called regions for the three clusters in the future, which give more information about the doings of the global forecast of the time series in the future.

Second step consists of getting a global forecast of the hourly global solar radiation time series using different parameters of the NAR network. Using the same number of clusters and embedding parameters, a phase space of this forecasted series was used. A comparison between the centroids of the global forecasted series and centroids of the regions can show the goodness of the forecast; the two centroids of each cluster region and forecasted series should be near each other.

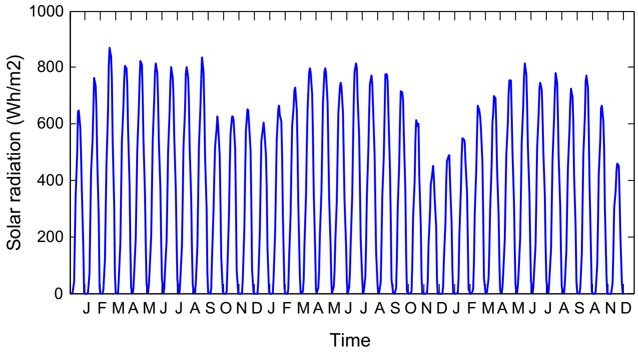


Fig. 4. Measured monthly global horizontal solar radiation data for Oran, Algeria.

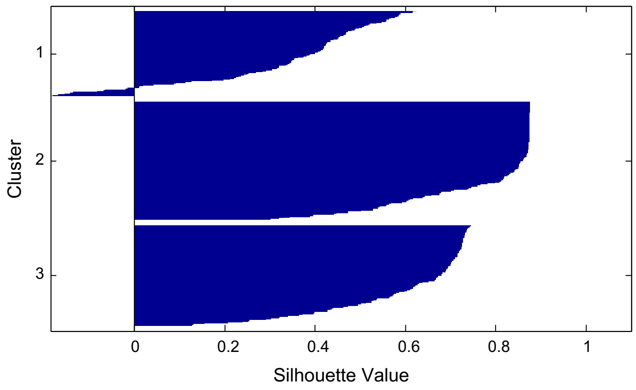


Fig. 5. Silhouette values with three clusters for the monthly horizontal solar radiation data for Oran, Algeria (from January 1994 to June 1996).

The results were shown in Fig. 6 which represents the phase space of the clustered regions presented by the sign of (+), and forecasted clusters represented by dots. According to Eq. (1), the plotted points represent the phase space of the solar radiation time series at time t and $t + 1$, that can visualize clearly the three kinds of clusters with low, medium and high solar radiation values. From this figure, it was shown that the most of the points of each forecasted cluster belonging to the appropriate regions. In addition, the centroids are close to each cluster that means that the forecasted series is quite good compared with the measured one.

Moreover, Fig. 7(a) shows the comparison results of the tested time series and forecasted one by the technique of k-means approach and NAR network and Fig. 7(b) represents the forecasted results using the ARMA model. The blue line is representing the testing monthly global horizontal solar radiation series between July 1996 to December 1996, and the red dot line is the forecasted series.

It's clearly shown that the forecasted series using the proposed method is virtually the same as the tested on with an RMSE equal to 60.24 W h/m^2 and NRMSE equal to 0.1985, which it was considered as a good forecast value compared to ARMA model that represents an NRMSE equal to 0.3078.

To understand more how this technique is working, we plotted a phase space of the hourly global horizontal solar radiation time

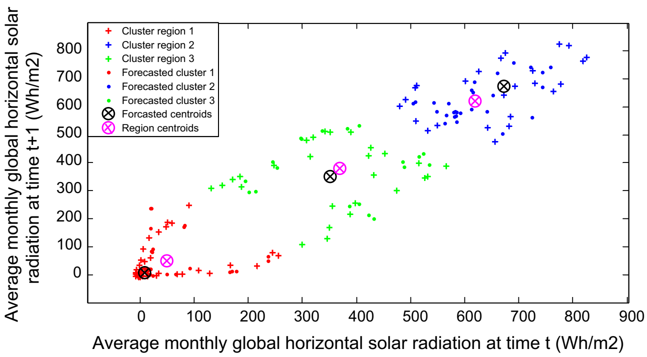


Fig. 6. Space phase reconstitution for forecasted regions and clusters of the monthly global horizontal solar radiation testing data (from July 1996 to December 1996).

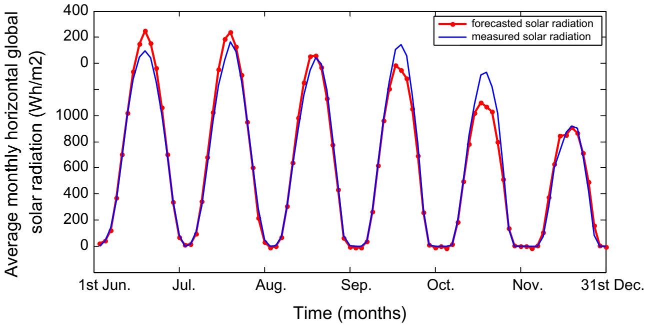


Fig. 7a. Comparison between measured monthly global horizontal solar radiation data (from July 1996 to December 1996), and forecasted by the proposed model.

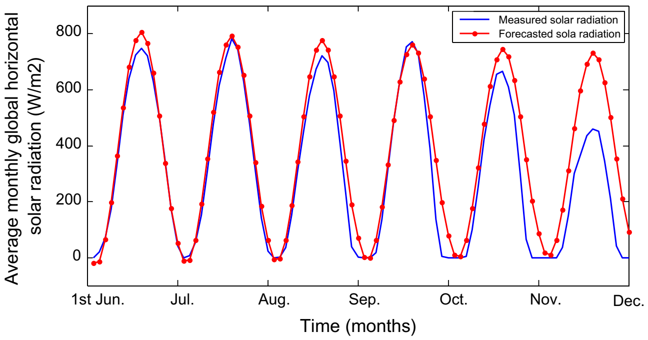


Fig. 7b. Comparison between measured monthly global horizontal solar radiation data (from July 1996 to December 1996), and forecasted by ARMA model.

series at time *t* and *t* + 1, but with wrong NAR network parameters, as shown in Fig. 8. It can observe from this figure that the points of the clusters are mixed with each other. In addition, the centroids are too far from each other, especially for the cluster 2 and 3, leading to the fact that the obtained forecast is not good comparing by the test one as shown in Fig. 9, which represented the forecasted average monthly global solar radiation series in red and the tested series in blue, representing an NRMSE error equals to 0.5532 that is not good forecast value.

In the same way, we used this methodology for more complicated solar radiation time series that provides forecasts at one-hour time step, which used widely in a lot of solar radiation application. Hence, an hourly global horizontal solar radiation time series for the year of 1996 was then applied.

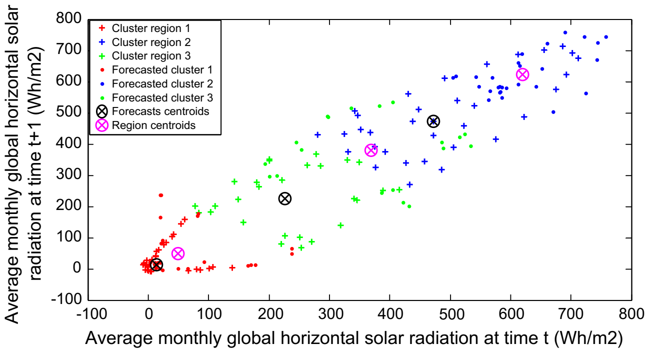


Fig. 8. Space phase reconstitution for forecasted regions and clusters of the monthly global horizontal solar radiation testing data (from July 1996 to December 1996) using wrong parameters.

The data were collected from the National Meteorological Office of Algeria for the site of Oran. We used only the data from sunrise to the sunset of the day. The data were divided into two sets, training set (from 1st January 1996 to 31st October 1996) that represent 4530 h, and test data set (from 1st November 1996 to the 31st December 1996) that represent 915 h. An example of one month from each season were shown in Figs. 10(a)–(d) that represent the hourly global horizontal solar radiation for months of January, April, July and October 1996 respectively.

A phase space reconstitution with time delay of 1 and embedding dimension equal to 2 were found experimentally using mutual information and false nearest neighbor methods to be the right choices for this time series. In addition, a plotting of the silhouette function with a different number of clusters was tested. We established that the use of three clusters to be the appropriate

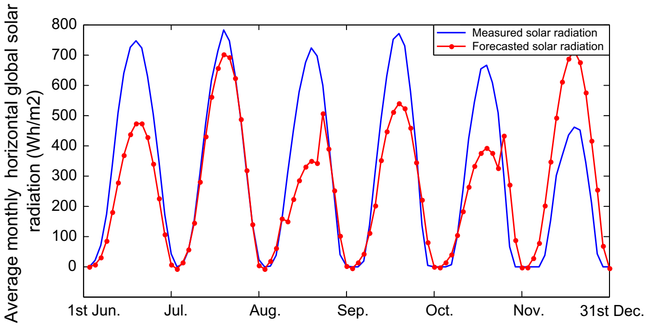


Fig. 9. Comparison between measured monthly global horizontal solar radiation data (from July 1996 to December 1996), and forecasted by proposed model using wrong parameters.

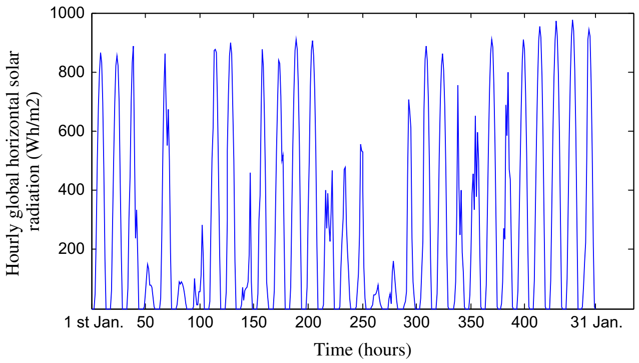


Fig. 10a. Measured hourly global horizontal solar radiation time series for January 1996.

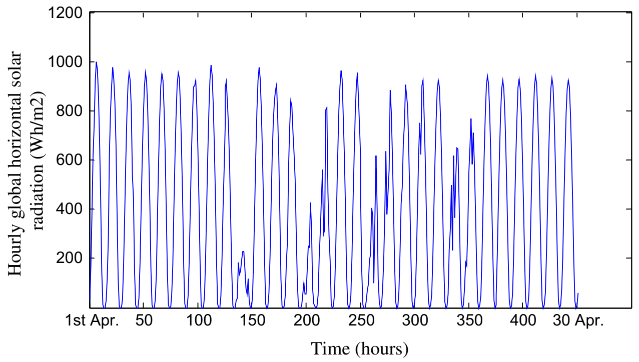


Fig. 10b. Measured hourly global horizontal solar radiation time series for April 1996.

choice as represented in Fig. 11, which represented the silhouette function of the hourly global horizontal solar radiation time series. It is clearly shown that the three clusters are well separated with the most of the points are above 0.6, except some negative ones in the second cluster that we can consider to be normal for such nonlinear time series.

For calculating the hourly global solar radiation time series, the *k*-means algorithm was then applied to clustering the training data. A local predictor was applied for obtaining future regions; those regions represent future windows for the forecasted data. Then, the NAR method with different time delay and neurons was applied to create global predictor of the data. The use of 25 delays with 13 neurons was found as the right choice for forecasting purpose. The results of phase space reconstitution of the forecasted regions and clusters for the hourly global horizontal solar radiation at time t and $t+1$ considering a time delay of 1 and embedding dimension of 2 was presented in Fig. 12.

From Fig. 12, the most of the points of each forecast cluster are in the right regions, the centroids are near each other, which mean that the obtained forecast is acceptable. The comparison between the forecasted hourly global horizontal solar radiation data and the tested data is shown in Fig. 13(a).

In addition, Figs. 13(b) and (c) represent the comparison results for the months of November 1996 and December 1996 respectively. The blue line represents measured data, and the red one is the forecasted data.

Moreover, the performance of the forecasted hourly global horizontal time series has been evaluated by calculating the RMSE errors between the actual data and forecasted one for the period of 1st November 1996 to 31st December 1996. The quadratic error between measured and simulated hourly global solar radiation using the proposed method was presented in Fig. 14. In addition, Fig. 15 represents the measured time series versus the forecasted time series.

From Fig. 14, the total RMSE was equal to 64.34 W h/m^2 and the NRMSE was 0.2003, which can be viewed as good forecasted values compared with an NRMSE equal to 0.3184 by using the baseline ARMA model.

In addition, from Fig. 15, the R squared value calculated by Eq. (12) is equal to 0.9330. The most of the points of the forecasted and measured series are near each to other. However, it presents some lags due to the total covered days that present a lot of clouds. Finally, from the simulation results, this methodology was conceived to be such a good method to perform the forecast results.

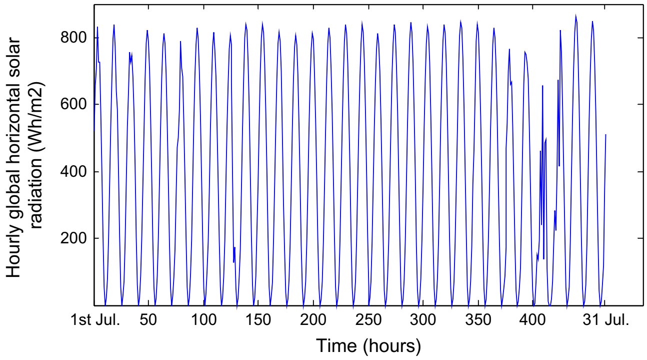


Fig. 10c. Measured hourly global horizontal solar radiation time series for July 1996.

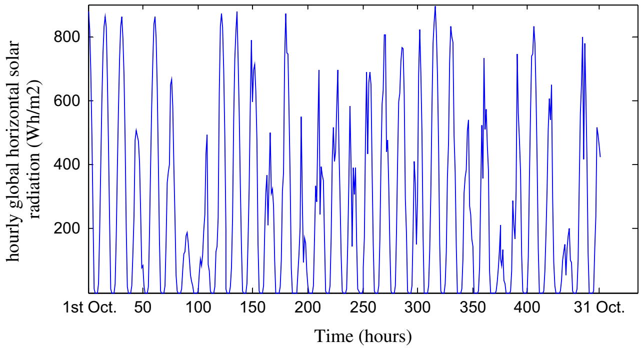


Fig. 10d. Measured hourly global horizontal solar radiation time series for October 1996.

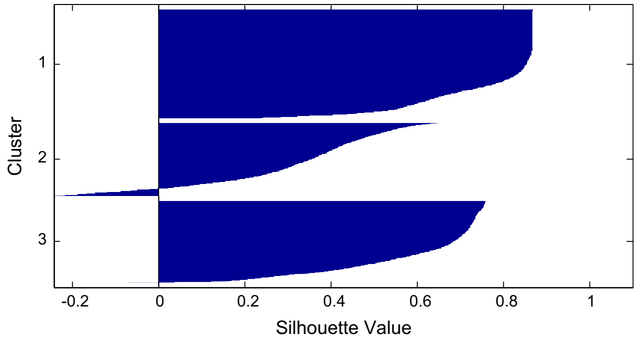


Fig. 11. Silhouette values with 3 clusters for the hourly global horizontal solar radiation data for Oran, Algeria (from 1st January 1996 to 31st October 1996).

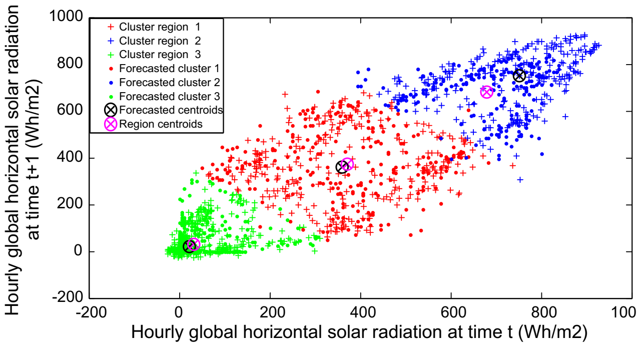


Fig. 12. Space phase reconstitution for the forecasted regions and clusters of the hourly global horizontal solar radiation testing data (from 1st November 1996 to the 31st December 1996).

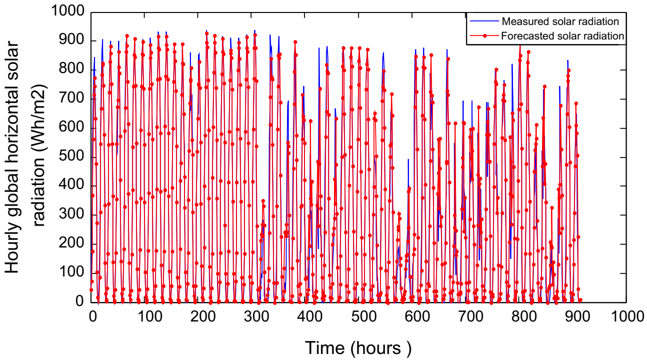


Fig. 13a. Comparison between measured hourly global horizontal solar radiation (from 1st November 1996 to the 31st December 1996) and forecasted by the proposed model.

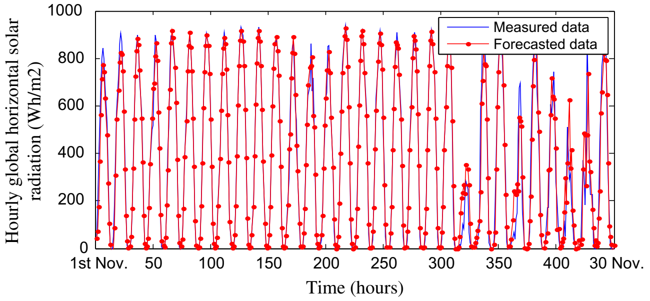


Fig. 13b. Comparison between measured hourly global horizontal solar radiation (from 1st November 1996 to the 31st November 1996) and forecasted by the proposed model.

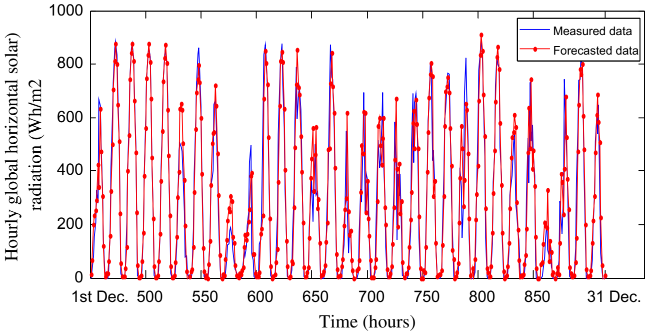


Fig. 13c. Comparison between measured hourly global horizontal solar radiation (from 1st December 1996 to the 31st December 1996) and forecasted by the proposed model.

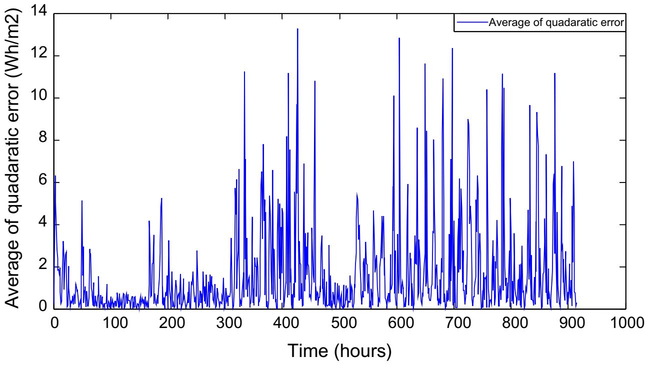


Fig. 14. the quadratic error between measured global horizontal solar radiation (from 1st November to 31st December 1996) and the forecasted using the proposed model.

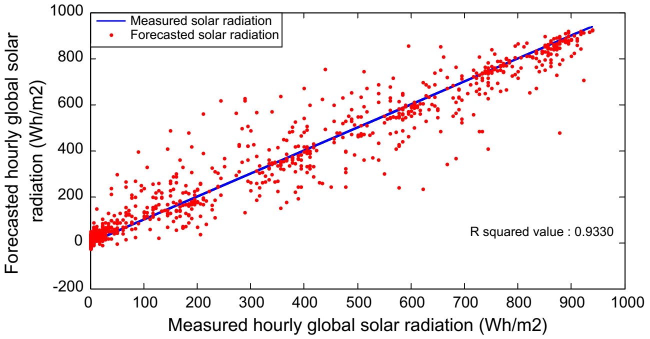


Fig. 15. The measured hourly global horizontal from (1st November 1996 to 31st December 1996) versus forecasted time series using the proposed model.

$$R^2 = \left(\frac{\left[\sum_{i=1}^{n} \left(I_{i,measured} - \bar{I}_{i,measured} \right) \right]^2}{\left[\sum_{i=1}^{n} \left(I_{i,predicted} - \bar{I}_{i,measured} \right) \right]^2} \right) \tag{12}$$

4. Conclusion

In this paper, we presented a time series forecasting methodology based on the clustering methods and artificial neural networks. The methodology consists of three essential stages. First, phase space reconstitution of the hourly global solar radiation time series was reached by finding the appropriate time delay using mutual information method, and the minimum embedding dimension is defined using false nearest neighbor method. Secondly, the unsupervised *k*-means clustering algorithm was then applied for grouping the input data into *k* clusters, which have similar characteristics. For choosing of the right number of clusters, the silhouette plot which represents a graphical representation of the separation of the heads of each cluster from another one was then used. Subsequently, a different NAR neural network was prepared on each cluster to act as a local predictor for the corresponding subspace of the input space. In addition, another NAR network was used to act as a global predictor for the solar radiation time series. The methodology was applied to generate multi-step ahead forecasts for the hourly global horizontal solar radiation time series. The obtained experimental results showed that the clustering of the input space is an important task to interpret the behavior of the series. Moreover, identifying forecasted regions using NAR network provides additional information about future patterns that can simplify the analysis of the global forecast of the series. In addition, the proposed method does not need a complicate clustering algorithm. Hence, as a conclusion of this work, the time series data mining method was considered such a good way of forecasting such similar problems. Nevertheless, this method presents some limitations for the total covered sky where the presence of clouds is heavy, also the calculation time, especially in the preparation phase of the NAR network. Hence, future works, will be focused on testing different clustering algorithms and different artificial neural networks to improve the forecasting performance that improves the reliability in the case of covered sky.

Acknowledgement

The authors would like to thank the University of Laghouat for the financial aspect of the present work.

References

[1] Maia André Luis S, de Carvalho ATFrancisco, Teresa BL. Forecasting models for interval-valued time series. Neurocomputing 2008;71:3344–52.
[2] Azadeh A, Maghsoudi A, Sohrabkhani S. An integrated artificial neural networks approach for predicting global radiation. Energy Convers Manage 2009;50:1497–505.
[3] Box GEP, Jenkins G. Time series analysis. Holden-Day (San Francisco, CA): Forecasting and Control; 1970.
[4] Celik AN, Muneer T. Neural network based method for conversion of solar radiation data. Energy Convers Manage 2013;67:117–24.
[5] Chen SX, Gooi HB, Wang MQ. Solar radiation forecast based on fuzzy logic and neural networks. Renew Energy 2013;60:195–201.
[6] Chow TWS, Leung CT. Non-linear autoregressive integrated neural network model for short term load forecasting. IEE Proc Gen Trans Distribut 1996;143:500–6.
[7] Faraway JJ, Chatfield C. Time series forecasting with neural networks: a case study. Statistics Group Research Report 9506, University of Bath; 1995.
[8] Fraser M, Swinney L. Independent coordinates for strange attractors from mutual information. Phys Rev A 1986;33:1134–40.
[9] Fu T. A review on time series data mining. Eng Appl Artif Intell 2011;24:164–81.
[10] Gelman R. 2011 Renewable energy data book (Revised Book). Efficiency Renew Energy (EERE) 2013.

[11] Huang Y. Advances in artificial neural networks-methodological development and application. Algorithms 2009;2:973–1007.
[12] Huang J, Korolkiewicz M, Agrawal M, Boland J. Forecasting solar radiation on an hourly time scale using a Coupled AutoRegressive and Dynamical System (CARDS) model. Sol Energy 2013;87:136–49.
[13] Haykin S. Neural networks: a comprehensive foundation. 2nd ed. Prentice Hall; 1998.
[14] Kaplanis S. New methodologies to estimate the hourly global solar radiation; comparisons with existing models. Renew Energy 2006;31:781–90.
[15] Kennel MB, Brown R, Abarbanel HD. Determining embedding dimension for phase space reconstruction using a geometrical construction. Phys Rev A 1992;45(6):3403–11.
[16] Kim HS, Eykholt R, Salas JD. Nonlinear dynamics, delay times, and embedding windows. Physica D 1999;127:48–60.
[17] Kostylev V, Pavlovski A. Solar power forecasting performance – towards industry standards, In: Proceedings of the 1st international workshop on the integration of solar power into power systems, Aarhus, Denmark; 2011
[18] Kugiumtzis D. State space reconstruction parameters in the analysis of chaotic time series—the role of the time window length. Physica D 1996;95:13–28.
[19] Levenberg K. A method for the solution of certain problems in least squares. Q Appl Math 1944;5:164–8.
[20] Lewis CD. International and business forecasting methods. London: Butter-Worths; 1982.
[21] Liao S, Chu P, Hsiao P. Data mining techniques and applications – a decade review from 2000 to 2011. Expert Syst Appl 2012;39:11303–11.
[22] Ljung L. System identification: theory for the user. 2nd ed. Prentice Hall PTR; 1998.
[23] Lletí R, Ortiz MC, Sarabia LA, Sánchez MS. Selecting variables for *k*-means cluster analysis by using a genetic algorithm that optimises the silhouettes. Anal Chim Acta 2004;515:87–100.
[24] MacKay DJC. Bayesian interpolation. Neural Comput 1992;4(3):415–47.
[25] MacQueen JB. Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley symposium on mathematical statistics and probability 1. University of California Press; 1967. p. 281–297.
[26] Mandelbrot BB. How long is the coastline of Britain? Statistical self-similarity and fractional dimension. Science 1967;155:636.
[27] Markham IS, Rakes TR. The effect of sample size and variability of data on the comparative performance of artificial neural networks and regression. Comput Oper Res 1998;25:251–63.
[28] Mellit A, Kalogirou SA, Hontoria L, Shaari S. Artificial intelligence techniques for sizing photovoltaic systems: a review. Renew Sustain Energy Rev 2009;13(2):406–19.
[29] Moller MF. A scaled conjugate gradient algorithm for fast supervised learning. Neural Networks 1993;4:525–33.
[30] Pandit SM, Wu SM. Time series and system analysis, with applications; 1983.
[31] Ragulskis M, Lukoseviciute K. Non-uniform attractor embedding for time series forecasting by fuzzy inference systems. Neurocomputing 2009;72:2618–26.
[32] Rousseeuw J. Silhouettes: a graphical Aid to the interpretation and validation of cluster analysis. Comput Appl Math 1987;20:53–65.
[33] Sandberg I, Xu L. Uniform approximation of multidimensional myopic maps. IEEE Trans Circ Syst I: Fund Theory Appl 1997;44(6):477–85.
[34] Spath H. Cluster dissection and analysis: theory, Fortran programs, examples translated by J. Goldschmidt. New York: Halsted Press; 1985. 226 pp.
[35] Takens F. Detecting strange attractors in turbulence, Dynamical Systems and Turbulence In: Rand DA, Young LS. editors, Lecture notes in mathematics. vol. 898; 1981. p. 366-381.
[36] Wu J, Chan KC. Prediction of hourly solar radiation using a novel hybrid model of ARMA and TDNN. Sol Energy 2011;85:808–17.
[37] Xu R, Donald C. Survey of clustering algorithm. IEEE Trans Neural Networks 2005;16(3):46–51.
[38] Zeng Z, Yang H, Zhao R, Meng J. Nonlinear characteristics of observed solar radiation data. Sol Energy 2013;87:204–18.
[39] Zhang G. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 2003;50:159–75.
[40] Boland JW. Time series and statistical modelling of solar radiation, Recent Advances in Solar Radiation Modelling, Viorel Badescu (Ed.), Springer-Verlag 2008; 283-312.