INFORMATICA, 2014, Vol. 25, No. 2, 265–282


DOI: http://dx.doi.org/10.15388/Informatica.2014.14

# A Hybrid Regression System Based on Local Models for Solar Energy Prediction

Héctor QUINTIÁN*, Jose Luis CALVO-ROLLE, Emilio CORCHADO
*University of Salamanca, Salamanca, Spain*
*e-mail: hector.quintian@usal.es, jlcalvo@udc.es, escorchado@usal.es*

Received: April 2012; accepted: September 2013

**Abstract.** The aim of this study is to predict the energy generated by a solar thermal system. To achieve this, a hybrid intelligent system was developed based on local regression models with low complexity and high accuracy. Input data is divided into clusters by using a Self Organization Maps; a local model will then be created for each cluster. Different regression techniques were tested and the best one was chosen. The novel hybrid regression system based on local models is empirically verified with a real dataset obtained by the solar thermal system of a bioclimatic house.

**Key words:** hybrid system, clustering, local models, SOM, MLP, SVM, prediction solar energy.

## 1. Introduction

Renewable energies are becoming more relevant in to the quest toward achieving a greater degree of sustainability (Demirbas, 2005). Consequently, their uses in houses is becoming more and more frequent. In Spain, the use of solar thermal energy has a higher incidence, due largely to the constraint (Real Decree 1027/2007, 2008) that forces the installation of solar thermal systems in new houses. The thermal power generated by renewable systems conditions thermal spending in the other non-renewal sources installed. If the daily power curve generated by the solar thermal system is known beforehand, it is possible use the consumption curve to predict when the energy supplied by the system will be insufficient (Martín *et al.*, 2010). In this way, corrective actions can be established to minimize energy supply needs by other non-renewable energy systems.

The proposal of this novel study is to use a hybrid intelligent system to predict the power energy generated by a solar thermal system based on consumption and solar radiation. There are other works based on solar systems prediction (Varol *et al.*, 2008; Caner *et al.*, 2011; Esen *et al.*, 2009) and other proposals which use intelligent techniques in other application fields like cooling systems (Arahal *et al.*, 2009), dental milling processes (Vera *et al.*, 2008), steering of ships (Calvo-Rolle and Quintián, 2012), mobiles telephones (Norkevicius and Raskinis, 2008) and so on. This requires dealing with a large amount of information, which makes the use of machine learning methods (Corchado *et al.*, 2010, 2012; Stankevicius, 2001) one of the best options. These methods have been

*Corresponding author.

demonstrated to be powerful non-linear estimators applicable to many real-world problems (Arpornwichanop and Shomchoam, 2009). They have been applied successfully in pattern recognition problems (Ripley, 2008; Ramirez-Cortes *et al.*, 2011), control and identification of systems (Souza and Barreto, 2010; Xuan *et al.*, 2011), time series prediction (Martínez-Rego *et al.*, 2011; Cherif *et al.*, 2011), and so on.

There are two main trends for the design of machine learning prediction models (Martínez-Rego *et al.*, 2011):

- Global models: One model is generated based on all training data available, trying to minimize the error for all input patters. Different techniques can be used to generate the model, such as Artificial Neural Networks (ANN) (Bishop, 2006) and Least Square Support Vector Machine (LS-SVM) (Ye and Xiong, 2007).
- Local models: In this case, the input space is divided into subsets based on similar characteristics of the input data, for which a clustering or quantization algorithm, such as K-means (Lei *et al.*, 2011), Self-Organizing Map (SOM) (Cherif *et al.*, 2011; Ghaseminezhad and Karami, 2011), or neural gas (Qin and Suganthan, 2005) is often used. Local models were first developed by Nowlan *et al.* (1991) and Jacobs and Jordan (1994), and were considered promising techniques in time series prediction. In order to determine the power thermal curve generated by solar thermal system, a system based on local models was developed for this research. The main novelty is the combination of different regression techniques to obtain the best local regression model for each cluster generated. For the clustering phase, a system based on SOM was applied.

This study is organized as follows. In Section 2, the different models used in this research are presented in formal way. Section 3 describes the proposed novel hybrid model. Section 4 presents the physical system on which the hybrid model is implemented. Section 5 describes the application of the hybrid model on the physical system. In Section 6 the results obtained are presented. Finally conclusions and future work are presented.

## 2. Used Models

### 2.1. *Self-Organizing Map (SOM)*

SOM (Kohonen, 1990) was developed as a visualization tool for representing high dimensional data on a low dimensional display. It is also based on the use of unsupervised learning (Baruque and Corchado, 2010). Typically, the array of nodes is one or two-dimensional, with all nodes connected to the $N$ inputs by an $N$-dimensional weight vector. The self-organization process is commonly implemented as an iterative on-line algorithm, although a batch version also exists. An input vector ($x$) is presented to the network and the node of the network in which the weights ($W_i$) are closest (in terms of Euclidean distance) to $x$, is chosen:

$$c = \arg\min_i \left(\|x - W_i\|\right). \tag{1}$$

The weights of the winning node and the nodes close to it are then updated to move closer to the input vector. There is also a learning rate parameter that usually decreases as the training process progresses. The weight update rule for inputs is defined as follows:

$$\Delta W_i = \eta h_{ci}[x - W_i], \quad \forall i \in N^{(c)}, \tag{2}$$

where, $W_i$ is the weight vector associated with neuron $i$, $x$ is the input vector, and $h$ is the neighbourhood function.

### 2.2. *Multilayer Perceptron (MLP)*

One of the most used ANN is the MLP (Bishop, 2006). The MLP is a supervised learning model. It is composed of one input layer, one or more hidden layers, and one output layer, all of them composed of neurons and pondered connections between the neurons of each layer. The training of this network is based on an iterative adjust of its weights with the objective to minimize the error between the output predicted by the network and the real output. Applying the *Theorem of Universal Approximation* (Hornik *et al.*, 1989), it can demonstrate that only one hidden layer is needed to model a non-linear projection between input and output layer. An MLP with one hidden layer, can be written mathematically as follows:

$$y_k^p = F_k\left(\sum_{i=1}^{L} w_{ik} \cdot F_i\left(\sum_{j=1}^{N} w_{ji} \cdot x_j^p + b_i\right) + b_k\right), \tag{3}$$

where:

$F_k \rightarrow$ activation function of neurons of the output layer.
$w_{ik} \rightarrow$ weight vector of connections from neurons of the hidden layer to neurons of the output layer.
$b_k \rightarrow$ bias of neurons of the output layer.
$k \rightarrow$ number of neurons of the output layer.
$F_i \rightarrow$ activation function of neurons of the hidden layer.
$w_{ji} \rightarrow$ weight vector of connections from neurons of input layer to neurons of hidden layer.
$b_i \rightarrow$ bias of neurons of the hidden layer.
$i \rightarrow$ number of neurons of the hidden layer.
$x_j^p \rightarrow$ $p$-th input pattern.
$j \rightarrow$ number of neurons of the input layer (equals to dimension of the input data).
$y_k^p \rightarrow$ predicted output for the $p$-th input pattern.

### 2.3. *Support Vector Regression (SVR)*

SVR is a modification of the algorithm of the Support Vector Machines (SVM) (Cristianini and Shawe-Taylor, 2000) for classification. In SVR the basic idea is to map the data into a high-dimensional feature space $F$ via a non-linear mapping and to perform a linear regression in this space.

2.3.1. *Least Square Support Vector Machine (LS-SVM)*

Least Square formulations of SVM are called LS-SVM. In this approximation the solution is obtained by solving a system of linear equations, which is comparable to SVM in terms of generalization of performance (Ye and Xiong, 2007). The application of LS-SVM to regression is known as LS-SVR (Least Square Support Vector Regression). In LS-SVR, the $\epsilon$-*insensitive* loss function is replaced by a classical squared loss function, which constructs the Lagrangian by solving the linear Karush–Kuhn–Tucker (KKT) system:

$$\begin{bmatrix} 0 & I_n^T \\ I_n & K + \gamma_{-1} I \end{bmatrix} \begin{bmatrix} b_0 \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}, \tag{4}$$

where $I_n$ is a $[n \times 1]$ vector of ones, $T$ means transpose of a matrix or vector, $\gamma$ a weight vector, $b$ regression vector and $b_0$ is the model offset.

In LS-SVR, only 2 parameters $(\gamma, \sigma)$ are needed. Where $\sigma$ is the width of the used kernel (Yankun *et al.*, 2007).

## 2.4. *Local Models*

When the data are unevenly distributed or distributed in groups, the results obtained by applying machine learning algorithms can be improved if multiple local models are created, as opposed to creating only one global model (Martínez-Rego *et al.*, 2011).

Local models are high-quality models applied to small regions of the input space of a learning problem. The advantage of local models is that they are often much more interesting and understandable to the domain expert, as they can concisely describe single aspects of the data instead of describing everything at once as global models do. In local models, the final system is composed of a group of local experts trained with data corresponding to one of the regions of the input data. In other words, the goal function $f$ is estimated as the union of the goal functions of each local model:

$$f(x) = \bigcup_{i=0}^{M} \hat{f}_i(x), \tag{5}$$

where $\hat{f}_i(x)$ is the goal function for each of the $M$ local models.

In this study each local model represents different operating areas of the system where diverse behaviours are present. Different machine learning techniques were applied to obtain the best non-linear model for each local model. The use of local models is not new in machine learning and they have been used before in Morik and Boulicaut (2005), Pang and Kasabov (2004), Cho *et al.* (2007), Souza and Barreto (2010). The inside of local models can be differentiated according to two types: the *classical local models*, which use linear models, and the *extended local models*, which use nonlinear models (Seher-Weiss, 2011). This study is focused on the second type, using non-linear local models.

To determine the global structure of the model, different unsupervised classification techniques can be used, such as SOM, K-means, standard winner-take-all (WTA), competitive learning, frequency sensitive competitive learning (FSCL), Principal Components Analysis (PCA) and the fuzzy competitive learning (FCL), to mention just a few.

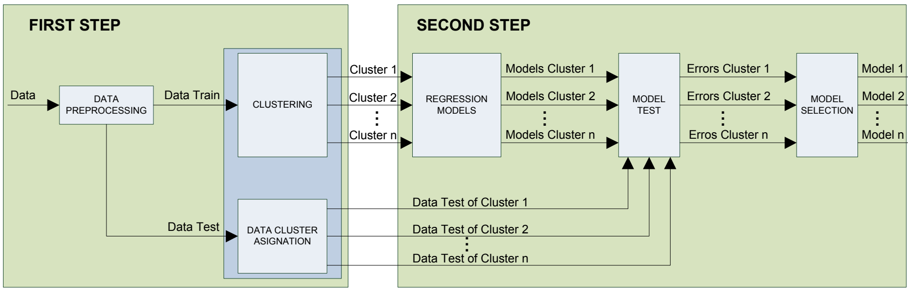



Fig. 1. Block diagram of the hybrid system.

After a comparison with other techniques, the SOM network was used to determine the number of clusters in the dataset for obtaining the best local model for each cluster.

## 3. Novel Hybrid Intelligent System

This research introduces a novel two-steps hybrid intelligent system to perform regression tasks with lower complexity and higher accuracy. The first step is to cluster data into sets of similar characteristics and the second is to develop regression models for each cluster, so that the best model for each cluster is selected (Fig. 1).

- **First step**:
  - **Preprocessing**: The original dataset is filtered and divided into two sets, one for training and the other for testing.
  - **Clustering**: Using the training dataset, an unsupervised learning system is trained for grouping data with similar characteristics into clusters. Then each sample of the training dataset is labelled with the cluster they belong to.
  - **Cluster assignment**: Each sample of the testing dataset is assigned to one of the generated models.
- **Second step**:
  - **Regression models**: Once the training data has been grouped into clusters, the next step is to generate regression models for each cluster by various techniques.
  - **Model test**: The regression models generated previously for each cluster are tested with the testing dataset in order to obtain the errors for each model.
  - **Model selection**: For each cluster, the different models generated are compared in terms of generated errors, and the best model for each cluster is then selected.

## 4. Physical System

The novel hybrid model described was applied for the prediction of thermal power values generated by a real physical system. It is a solar thermal collection system installed within a bioclimatic house. The physical system will now be described in detail.

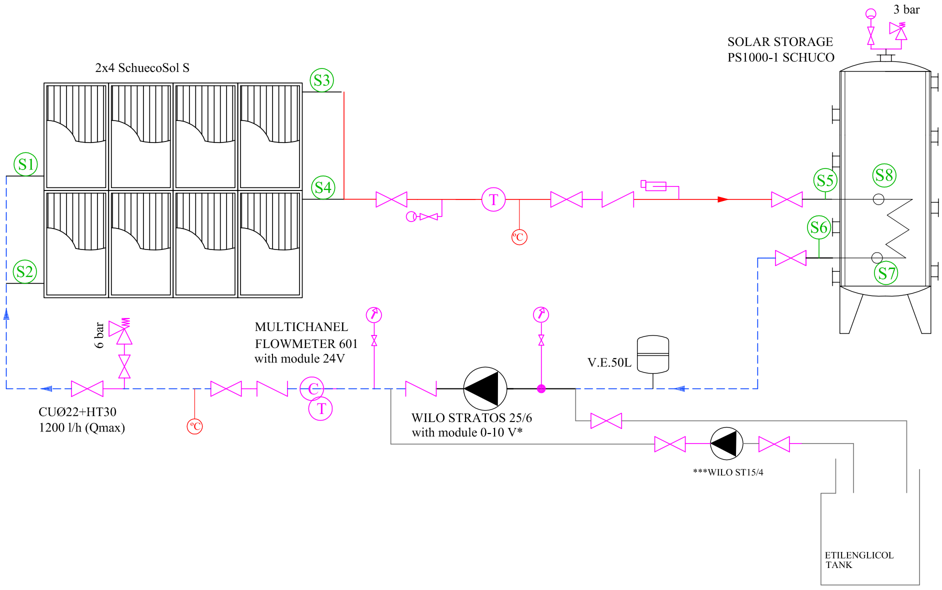



Fig. 2. Solar thermal system installed

### 4.1. *System Description*

Figure 2 shows the house solar thermal system, from the solar energy collection phase to the accumulation phase. The solid red line represents the part of the hydraulic circuit that transports hot fluid, while the dashed blue line represents the cold fluid. The collection zone of solar thermal system consists of solar panels (with an inclination of 19° in the North facade) that communicate with a solar accumulator through an ethyleneglycol closed hydraulic circuit. The hydraulic pump boosts the ethyleneglycol to solar panels entries $S1$ and $S2$, which are arranged in a structure of two parallel blocks with four panels in series per block. The ethyleneglycol circulates inside the solar panels, absorbs the heat and sends it out through $S3$ and $S4$ with a temperature higher than the input. From there, the fluid is driven to the entrance of the parallel plate heat exchanger of the solar accumulator ($S8$), where the energy stored in the ethyleneglycol heats the water stored inside the solar accumulator. Once the fluid leaves the accumulator ($S6$), the hydraulic circuit is closed to send the fluid back to solar panels, boosting it with a circulation pump.

### 4.2. *Energy Capacities of the System*

In the design phase, it has estimated (Sotavento, 2013) that the solar collection system covers 90.37% of the house HSW (Hot Sanitary Water) demand. The yield of the solar panels is 78.1%, with losses in a range from 5% to 10% due to the inclination of the solar panels, as the optimum inclination of the panels is 43° for the location of the house.

### 4.3. *Data Description for Model Generation*

This study used a dataset corresponding to one year of sampling at a bioclimatic house. This dataset was filtered by removing all outlayers, which produce a final dataset of 36,292 samples (12 months). Each sample has two inputs (*flow* in the solar thermal circuit and *solar radiation*) and one output (thermal *power* generated by the solar thermal system). The equipment used for measuring is:

- Power meter: Kamstrup type Multichannel 601, it can measure *thermal power*, *flow* and temperature.
- Radiation meter: Apogee model PYR-P, it can measure *solar radiation* with a sensitivity of 0.200 mV per $\mathrm{Wm}^{-2}$.

The time format of the data is UMT, with a ten minute measurement period. The data set period recorded goes since 1st September, 2010 until 31st August, 2011.

## 5. Hybrid System Applied to a Real Physical System

To determine the validity degree of the proposed novel hybrid system, it was applied to a real physical system using two regression techniques: LS-SVR and ANN, as shown in Fig. 3(b).

### 5.1. *Hybrid System Components*

Each of the blocks of the hybrid system introduced in Section 3 are now described in detail for their implementation into the physical system using two regression techniques within the hybrid system (LS-SVR and ANN).

#### 5.1.1. *Dataset Preprocess*

The original dataset is filtered and divided into two sets, one for training and the other for testing. Two-thirds of the filtered data are used for training (2205 samples, March 2011), while the remaining third is used for testing (1102 samples, March 2011).

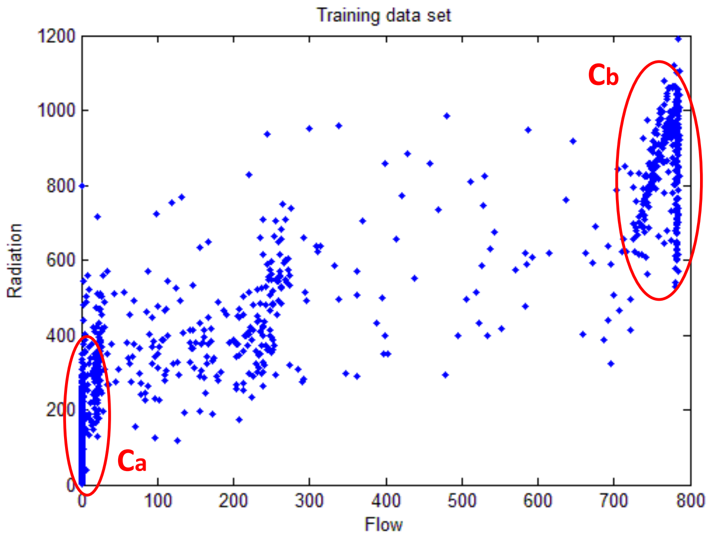
#### 5.1.2. *Clustering*

For clustering, 2 unsupervised learning techniques were tested: Principal Components Analysis (PCA) (Oja *et al.*, 1992) and SOM (Kohonen, 1990).

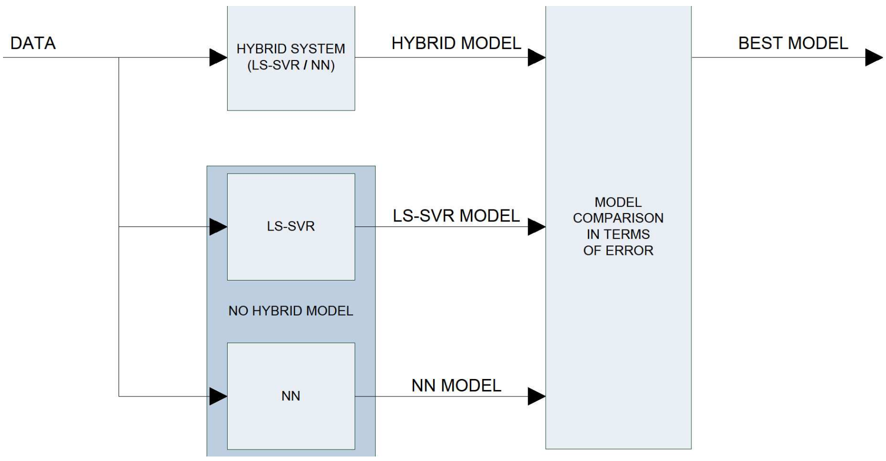
An initial visual analysis of the dataset is shown in Fig. 3(a). It clearly shows the existence of at least two distinct sets of related data labelled Ca and Cb (Fig. 3(a)).

Figure 4 shows the results of these two techniques. PCA only confirms the existence of the initial clusters (Ca and Cb, Fig. 3(a)). Since SOM shows the existence of one more cluster, it was selected as the clustering technique.

SOM toolbox for Matlab (Vesanto *et al.*, 2000) was used. The SOM network was trained using a grid size of $20 \times 20$ units, a gaussian neighborhood function with $\sigma_{ini} = 5$ decreasing linearly to $\sigma_{fin} = 1$, linear learning rate function, using 2000 iterations.




(a) Dataset

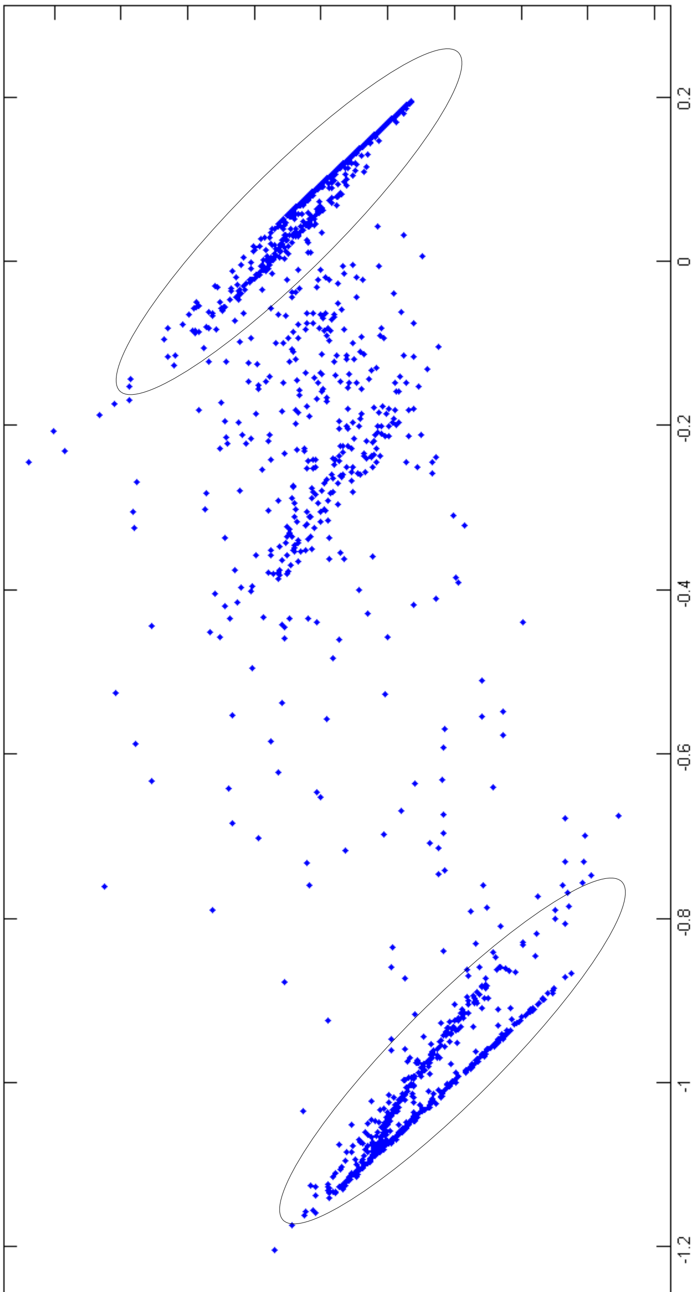



(b) Novel hybrid system components

Fig. 3. Dataset with initial cluster Ca and Cb (a) and components of the novel hybrid system (b).

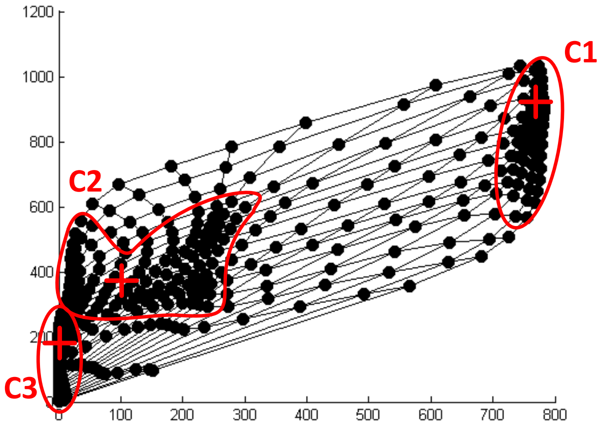
Figure 4(a) shows the map after training, which includes a high density of near nodes in three different parts of the map. It confirms the existence of the initial clusters (Ca and Cb, Fig. 3(a)) and it shows the existence of a third cluster. Three clusters were marked in the final map (C1, C2 and C3 Fig. 4(a)) as well as the representative node of each cluster (crosses in Fig. 4(a)).

Once the SOM model is obtained, each sample is labelled according to the representative node to which it belongs, using the criterion of least Euclidean distance to each representative node. Separation is obtained for the initial set of data into subsets (clusters) of related data. Figures 5(a), 5(b) and 5(c), show the distribution of the samples of the training dataset in each cluster.




(a) PCA

Fig. 4. Results for PCA and SOM.



(b) SOM

Fig. 4. (*Continued*.)

The testing data are also labelled according to the least Euclidean distance. When they reach to the block *Model test*, they will be used to test the models.

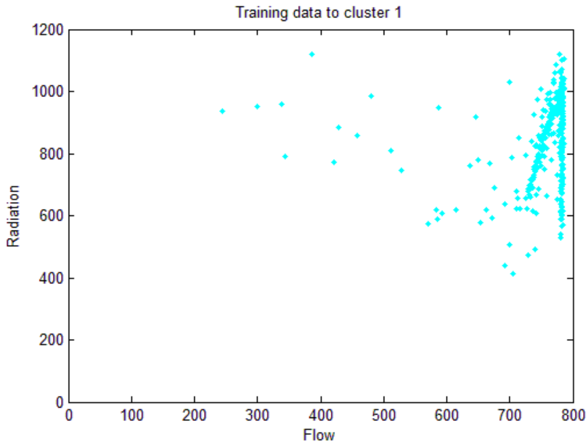
### 5.1.3. *Regression Models*

This block aims to generate regression models based on different techniques, such as SVM, ANN, etc. Each regression technique is trained with data from each cluster, and the output of the block will have $m \times n$ models, where $m$ is the number of regression techniques and $n$ is the number of clusters.

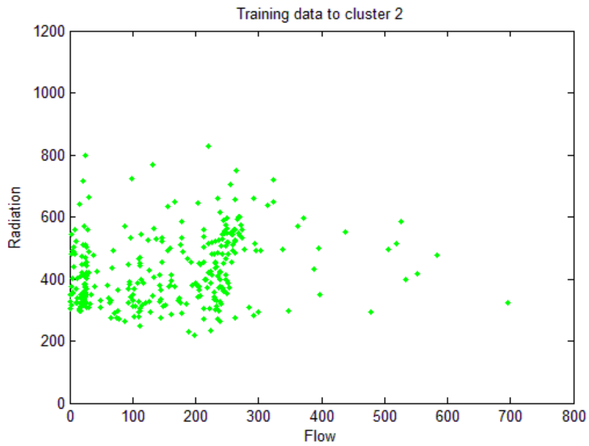
Two regression techniques were used in this study:

- ANN: MLP was used. It was tested with 5, 10 and 15 neurons in the hidden layer, log sigmoidal and tangent sigmoidal transfer function for the hidden layer and linear for the output layer. To train the MLP, the Levenberg–Marquardt optimization algorithm (Levenberg, 1944) was used to update the weights and bias of the network, as it is often faster than classical error backpropagation algorithm (Matworks Documentation Center, 2013). Finally the best results were obtained using 10 neurons and a sigmoidal transfer function for the hidden layer.
- SVM: LS-SVM (Least Square Support Vector Machine) (De Brabanter *et al.*, 2010) Matlab toolbox was used. In this toolbox, the tuning of the parameters $\gamma$, $\sigma$ (4) is conducted in two steps. First, a state-of-the-art global optimization technique, Coupled Simulated Annealing (CSA) (Xavier de Souza *et al.*, 2010), determines suitable parameters according to specific criterion. These parameters are then given to a second optimization procedure (simplex or gridsearch) to perform a fine-tuning step.

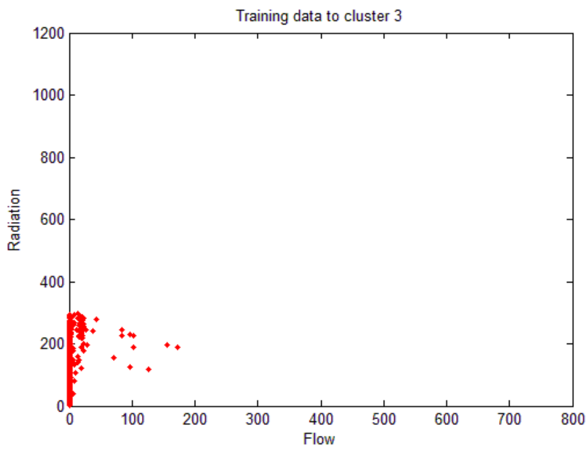
A cross validation of 10 folds (Kohavi, 1995) were used in the training of all regression techniques.




(a) Cluster 1 dataset




(b) Cluster 2 dataset




(c) Cluster 3 dataset

Fig. 5. Data clusters 1(a), 2(b) and 3 (c).

Table 1
Total number of samples for training and testing for each month before and after clustering.

| Month | Total | | Cluster 1 | | Cluster 2 | | Cluster 3 | |
|---|---|---|---|---|---|---|---|---|
| | Train | Test | Train | Test | Train | Test | Train | Test |
| January (2011) | 2221 | 1111 | 1571 | 776 | 250 | 144 | 400 | 191 |
| February (2011) | 2205 | 1102 | 1568 | 769 | 211 | 104 | 426 | 229 |
| March (2011) | 2230 | 1115 | 1458 | 733 | 236 | 130 | 536 | 252 |
| April (2011) | 1819 | 909 | 948 | 453 | 206 | 126 | 665 | 330 |
| May (2011) | 2012 | 1006 | 1107 | 560 | 268 | 131 | 637 | 315 |
| June (2011) | 2040 | 1021 | 1189 | 600 | 260 | 136 | 591 | 285 |
| July (2011) | 2000 | 1000 | 1229 | 613 | 245 | 116 | 526 | 271 |
| August (2011) | 2001 | 1000 | 1211 | 604 | 237 | 133 | 553 | 263 |
| September (2010) | 2880 | 1440 | 1940 | 971 | 281 | 144 | 659 | 325 |
| October (2010) | 1400 | 700 | 817 | 412 | 185 | 73 | 398 | 215 |
| November (2010) | 1384 | 693 | 905 | 459 | 189 | 96 | 290 | 138 |
| December (2010) | 2002 | 1001 | 1414 | 722 | 232 | 93 | 356 | 186 |

#### 5.1.4. *Model Test*

Once all models are trained, each model is tested against test data (1/3 of total), determining the NMSE (Normalized Mean Square Error) for each model.

#### 5.1.5. *Model Selection*

In this block, each of the previously trained models are selected based on their NMSE, thus obtaining the best model for each cluster. The NMSE of the selected models is compared to the NMSE of each non hybrid regression model.

## 6. Results

In this section we compare the results obtained by the novel hybrid system and other non-hybrid system in terms of NMSE. For both hybrid and non-hybrid systems, ANN and SMV regression techniques are used to create regression models. Table 1 provides information on the total data size for each month and the separation of the data into training and testing datasets both before (shown in the Total column) and after the clustering process.

### 6.1. *Results of the Hybrid Model*

The results of the hybrid model are shown in Table 2. Table 2 shows the NMSE of each model generated by the hybrid system for each cluster.

### 6.2. *Results of Non-Hybrid Models*

After obtaining the results of the hybrid model, it is compared to the non-hybrid model. The following NMSE results for non-hybrid model are obtained (Table 3):

Comparing the results of Tables 2 and 3, it can see that the novel hybrid model proposed always obtains better results than the non-hybrid models.

Using only local models and one regression technique (columns *Total* of ANN or LS-SVR in Table 2), the results are better than non-hybrid models, but in 7 of the 12 months,

Table 2
Values of NMSE of the hybrid system.

| Month | Hybrid | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | LS-SVR | | | | ANN | | | | Combined |
| | C1 | C2 | C3 | Total | C1 | C2 | C3 | Total | Total |
| January | 0.0146 | 0.4211 | 0.0870 | 0.0494 | 0.0102 | 0.5184 | 0.0648 | **0.0369** | **0.0358** |
| February | 0.3973 | 0.3501 | 0.0796 | **0.0316** | 0.6518 | 0.3810 | 0.0813 | 0.0323 | 0.0316 |
| March | 0.3090 | 0.1961 | 0.0733 | 0.0230 | 0.2716 | 0.2160 | 0.0612 | **0.0199** | **0.0191** |
| April | 0.6294 | 0.0540 | 0.0120 | **0.0113** | 0.8663 | 0.1649 | 0.0235 | 0.0246 | 0.0113 |
| May | 0.2962 | 0.3278 | 0.2352 | **0.0601** | 0.4166 | 0.4083 | 0.2370 | 0.0621 | 0.0601 |
| June | 0.1770 | 0.3144 | 0.1954 | **0.0508** | 0.7445 | 0.3387 | 0.1924 | 0.0517 | **0.0501** |
| July | 0.3196 | 0.3810 | 0.1937 | **0.0466** | 0.4013 | 0.4185 | 0.2199 | 0.0527 | 0.0466 |
| August | 0.1739 | 0.3907 | 0.1741 | **0.0393** | 0.7408 | 0.3667 | 0.1977 | 0.0441 | **0.0390** |
| September | 0.7069 | 0.2751 | 0.0948 | **0.0364** | 0.8510 | 0.2620 | 0.0981 | 0.0376 | **0.0363** |
| October | 0.8036 | 0.1799 | 0.1391 | **0.0319** | 0.8341 | 0.2300 | 0.1443 | 0.0359 | 0.0319 |
| November | 0.2013 | 0.3809 | 0.1441 | **0.0355** | 0.8608 | 0.4082 | 0.1364 | 0.0447 | **0.0322** |
| December | 0.5034 | 0.6418 | 0.1654 | **0.0510** | 0.1649 | 0.6032 | 0.2017 | 0.0627 | **0.0489** |

Table 3
NMSE of non-hybrid system.

| Non-hybrid | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Month | Jan. | Feb. | Mar. | Apr. | May | Jun. | Jul. | Aug. | Sep. | Oct. | Nov. | Dec. |
| LS-SVR | 0.129 | 0.159 | 0.116 | 0.115 | 0.194 | 0.211 | 0.187 | 0.183 | 0.156 | 0.105 | 0.120 | 0.219 |
| ANN | 0.181 | 0.203 | 0.128 | 0.098 | 0.299 | 0.246 | 0.269 | 0.187 | 0.168 | 0.134 | 0.083 | 0.227 |

when combining the best results of both techniques, the global results (column *Combined*, Table 2) are improved.

Focusing on the results for each cluster, the LS-SVR technique obtains better results more times than ANN. This is also the case in the non-hybrid models (Table 3).

Table 2 shows that cluster 3 gets the best results of all the clusters, while clusters 2 and 1 have similar results. Looking at Figs. 5(a)–5(d), it is possible to determine that for cluster 3, only a few samples fell outside the limits. However, for the other two clusters there are several samples outside of their limits, so that the errors for cluster 3 are lower than those of the others two clusters.

Finally, Fig. 6 shows the fit between real output (dashed blue line) and the hybrid model predicted output (solid red line) for the testing dataset for October, where the vertical axis represents the power of the solar thermal system and the horizontal axis represents the number of samples for this month. Figure 6 shows a zoom over the selection of Fig. 6. Figure 6 shows that the prediction of the proposed novel hybrid model is very similar to the real data, even in those parts of the graph where sudden changes in values occur (graphic peaks).

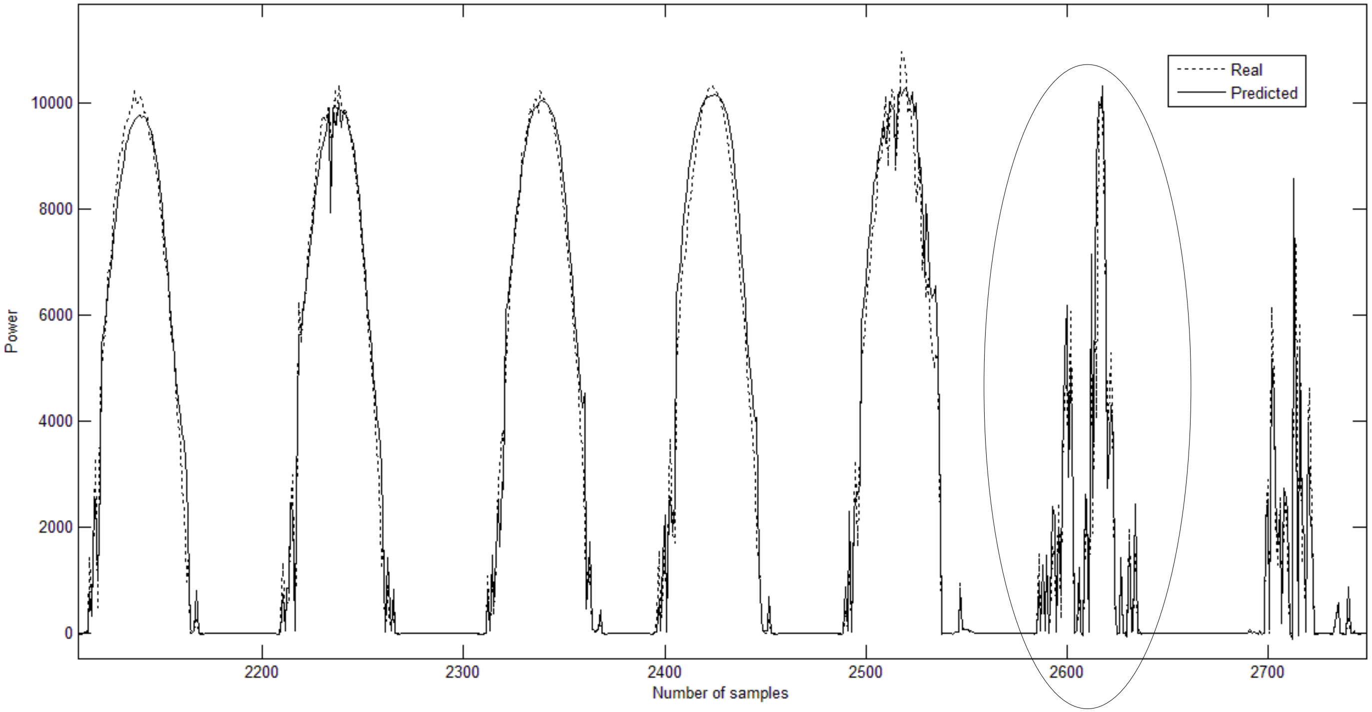



Fig. 6. Graph of the fit between real output and predicted hybrid model output for the testing dataset for October.

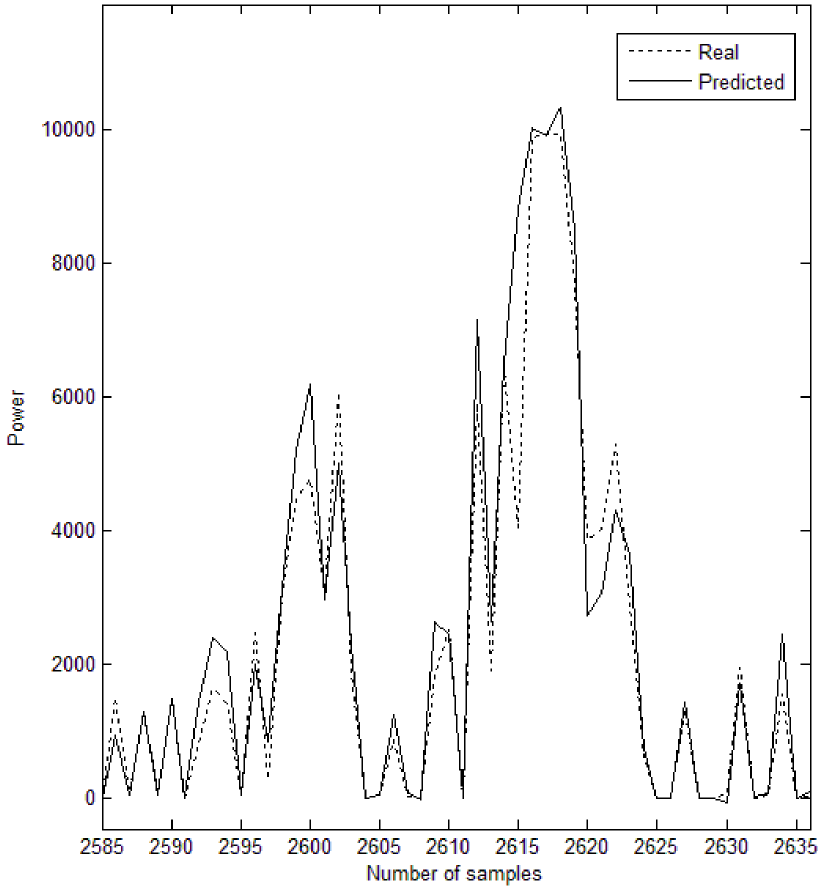



Fig. 7. Zoom on the selected part of Fig. 6.

## 7. Conclusions and Future Work

This research presents a novel hybrid regression system which was then analysed and compared to other models. This hybrid system is based on local models combined with intelligent techniques for regression. The main advantage of the proposal is the high accuracy combined with the low complexity of the system. The present study also analyses the application of the novel hybrid system on a real case based on a solar thermal system installed in a bioclimatic house. The achieved results show that the proposed hybrid system makes possible to predict the thermal power generated by solar thermal system, which provides a much better approach than the non-hybrid systems. Future works will be focused on different regression techniques, other unsupervised learning techniques, and other case studies, especially when the system to approximate has a non-linear component and the operation points are very differentiated, as with systems of level control, temperature, pressure, and so on.

**Acknowledgements**. This research is partially supported by the Spanish Ministry of Economy and Competitiveness under project TIN2010-21272-C02-01 (funded by the European Regional Development Fund), SA405A12-2 from Junta de Castilla y León. We would like to thank the Instituto Enerxético de Galicia (INEGA) and Parque Eólico Experimental de Sotavento for its technical support in this work.

## References

Arahal, M.R., Amparo, N.R., Ignacio, A.A., Francisco, R. (2009). Local model sets with internal description. Application to a cooling solar plant. *RIAI – Revista Iberoamericana de Automatica e Informatica Industrial*, 6(1), 51–62.

Arpornwichanop, A., Shomchoam, N. (2009). Control of fed-batch bioreactors by a hybrid on-line optimal control strategy and neural network estimator. *Neurocomputing*, 72, 2297–2302.

Baruque, B., Corchado, E. (2010). A weighted voting summarization of SOM ensembles. *Data Mining and Knowledge Discovery*, 21(3), 398-426.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Calvo-Rolle, J.L., Quintián, H. (2012) Advances in Expert Systems: Neuro-Knowledge Model Based on a PID Controller to Automatic Steering of Ships. *Intech*, ISBN 980-953-307-510-0.

Caner, M., Gedik, E., Keēebas, A. (2011). Investigation on thermal performance calculation of two type solar air collectors using artificial neural network. *Expert Systems with Applications*, 38(3), 1668–1674.

Cherif, A., Cardot, H., Boné, R. (2011). SOM time series clustering and prediction with recurrent neural networks. *Neurocomputing*, 74, 1936–1944.

Cho, J., Principe, J., Erdogmus, D., Motter, M. (2007). Quasi-sliding mode control strategy based on multiple linear models. *Neurocomputing*, 70(4–6), 962–974.

Corchado, E., Abraham, A., Carvalho Demirbas, A. (2010). Hybrid intelligent algorithms and applications. *Information Sciences*, 180, 2633–2634.

Corchado, E., Graña, M., Woźniak, M. (2012). New trends and applications on hybrid artificial intelligence systems. *Neurocomputing*, 75(1), 61–63.

Cristianini, N., Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge.

De Brabanter, K., Karsmakers, P., Ojeda, F., Alzate, C., De Brabanter, J., Pelckmans, K., De Moor, B., Vandewalle, J., Suykens, J.A.K. (2010). LS-SVMlab toolbox user's guide version 1.7. `http://www.esat.kuleuven.be/sista/lssvmlab/`.

Demirbas, A. (2005). Potential applications of renewable energy sources, biomass combustion problems in boiler power systems and combustion related environmental issues. *Progress in Energy and Combustion Science*, 31, 171–192.

Esen, H., Ozgen, F., Esen, M., Sengur, A. (2009). Modelling of a new solar air heater through least-squares support vector machines. *Expert Systems with Applications*, 36(7), 10673–10682.

Ghaseminezhad, M.H., Karami, A. (2011). A novel self-organizing map (SOM) neural network for discrete groups of data clustering. *Applied Soft Computing*, 11, 3771–3778.

Hornik, K., Stinchcombre, M., White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks* 2, 359–366.

Jacobs, R.A., Jordan, M.I. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6, 181–214.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *14th International Joint Conference on Artificial Intelligence*, pp. 1137–1145.

Kohonen, T. (1990). *The self-organizing map*. Proceedings of the IEEE, 78(9), 1464–1480.

Lei, H., Long, L.R., Antani, S., Thoma, G.R. (2011). Multiphase level set model with local K-means energy for histology image segmentation. In: *2011 First IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology (HISB)*, pp. 32–39.

Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 5, 164–168.

Martín, L., Zarzalejo, L.F., Polo, J., Navarro, A., Marchante, R., Cony, M. (2010). Prediction of global solar irradiance based on time series analysis: application to solar thermal power plants energy production planning. *Solar Energy*, 84(10), 1772–1781.

Martínez-Rego, D., Fontenla-Romero, O., Alonso-Betanzos, A. (2011). Efficiency of local models ensembles for time series prediction. *Expert Systems with Applications*, 38, 6884–6894.

Matworks Documentation Center (02/01/2013). Levenberg–Marquardt backpropagation. `http://www.mathworks.es/es/help/nnet/ref/trainlm.html`.

Morik, K., Boulicaut, J.F. (2005). *Local Pattern Detection and Clustering*. *Lecture Notes in Computer Science*, Vol. 3539. Springer, Berlin, pp. 53–70.

Norkevicius, G., Raskinis, G. (2008). Modeling phone duration of Lithuanian by classification and regression trees, using very large speech corpus. *Informatica*, 2(19), 271–284.

Nowlan, S.J., Hinton, G.E., Jacobs, R.A., Jordan, M.I. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3, 79–87.

Oja, E., Ogawa, H., Wangviwattana, J. (1992). Principal components analysis by homogeneous neural networks, Part 1. The weighted subspace criterion. *IEICE Transaction on Information and Systems*, E75D, 366–375.

Pang, S., Kasabov, N. (2004). Inductive vs transductive inference, global vs local models: SVM, TSVM, and SVMT for gene expression classification problems. *International Joint Conference on Neural Networks*, 2, 1197–1202.

Qin, A.K., Suganthan, P.N. (2005). Enhanced neural gas network for prototype-based clustering. *Pattern Recognition*, 38, 1275–1288.

Ramirez-Cortes, J.M., Gomez-Gil, Alarcon-Aquino, P.V., Baez-Lopez, D., Enriquez-Caldera, R. (2011). A biometric system based on neural networks and SVM using morphological feature extraction from hand-shape images. *Informatica*, 22(2), 225–240.

Real Decree 1027/2007 of 20 July, which approves the Regulation Thermal Installations in Buildings (BOE No. 207 of 29.08.2007).

Ripley, B.D. (2008). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Seher-Weiss, S. (2011). Identification of nonlinear aerodynamic derivatives using classical and extended local model networks. *Aerospace Science and Technology*, 15, 33–44.

Souza, L.G.M., Barreto, G.A. (2010). On building local models for inverse system identification with vector quantization algorithms. *Neurocomputing*, 73(10–12), 1993–2005.

Sotavento online documentation (02/01/2013). *Bioclimatic House Complete Memory*. `http://www.sotaventogalicia.com/pdf/viv_mem_completa.pdf`.

Stankevicius, G. (2001). Forming of the investment portfolio using the self-organizing maps (SOM). *Informatica*12 (4), 573–584.

Varol, Y., Koca, A., Öztop, H.F., Engin Avci, D.H. (2010). Forecasting of thermal energy storage performance of phase change material in a solar collector using soft computing techniques. *Expert Systems with Applications*, 37, 2724–2732.

Vera, V., Corchado, E., Redondo, R., Sedano, J., García, Á. E. (2012). Applying soft computing techniques to optimise a dental milling process. *Neurocomputing*. `http://dx.doi.org/10.1016/j.neucom.2012.04.033`.

Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J. (2000). *SOM toolbox for Matlab 5*. `http://www.cis.hut.fi/somtoolbox/`.

Xavier de Souza, S., Suykens, J.A.K., Vandewalle, J., Bolle, D. (2010). Coupled simulated annealing. *IEEE Transactions on Systems, Man and Cybernetics – Part B*, 40(2), 320–335.

Xuan, H., Wen-Fang, X., Zhijun, F., Weidong, L. (2011). Nonlinear systems identification using dynamic multi-time scale neural networks. *Neurocomputing*, 74, 3428–3439.

Yankun, L., Xueguang, S., Wensheng, C. (2007). A consensus least support vector regression (LS-SVR) for analysis of near-infrared spectra of plant samples. *Talanta*, 72, 217–222.

Ye, J., Xiong, T. (2007). SVM versus least squares SVM. In: *The 11th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 640–647.

**H. Qintián** He graduated as Industrial Engineer in 2010, and received the MSc degree in Computer Science by University of Coruña, Spain in 2011. He has a predoctoral grant to develop his PhD studies in Hybrid Intelligent Systems. He has published more than 20 papers in prestigious journals and conferences proceedings. He has been editor of prestigious international conferences (HAIS, SOCO, CISIS). His research interests include neural networks, with a particular focus on exploratory projection pursuit, maximum likelihood hebbian learning, self-organising maps, feature selection and optimization problems and its applications to industrial and environmental problems.

**J.L. Calvo-Rolle** was born in Coruña, Spain, in 1974. He received the MS and PhD degrees in Industrial Engineering from the University of Leon, Leon, Spain, in 2004, and 2007, respectively. He is Associate Professor of Automatic Control and the head of Industrial Engineering Department, Faculty of Engineering, University of Coruña, Spain. His main research interests have been centered in applying expert system technology to the diagnosis and control systems and in intelligent training systems for control engineering, optimization and education.

**E. Corchado** (Accredited as Full Professor by ANECA – The National Agency for Quality Assessment and Accreditation of Spain – 15th February 2013). He has been elected as the 2014–2015 IEEE Spanish Section President-Elect. He is Vice-President of IEEE Spain Section (2012–2013), chair of the IEEE Systems, Man and Cybernetics-Spanish Chapter during 2009-12; and was member of the IEEE R8 PA subcommittee during 2011–12. He is co-author of several patents and Intellectual properties. He has been reviewer an evaluator for the European Commission since 2007. He has senior scientific and managerial leadership in the area of software engineering, CI and quality assurance, an extended record of more than 60 publications in high reputed scientific journals and also acting as member of Editorial Boards from several reputed international journals (KAIS/Springer, Computational Intelligence/WILEY, IJCIA/World Scientific, and IJRIS/INDERSCIENCE). He is guest editor of several "Special Issues" of reputed journals as Information Science (Elsevier), Knowledge and Information Systems (KAIS, Springer), Neurocomputing (Elsevier), International Journal of Neural Systems (IJNS, World Scientific), Logic Journal of the IGPL (Oxford Journals), Neural Networks World. He is/was General Chair/Scientific Chair of several reputed conferences (HAIS, IDEAL, SOCO, CISIS, NABIC, CASON, NWESP, WICT). He has participated on several European Projects (ENGINE, SOFT-COMP, EMaCS) and Cost Actions (IntelliCIS (FP7), NISIS (FP6), AAPELE) and is member of a EU Centre of Excellence (IT4Innovation). His research interests include neural networks, with a particular focus on exploratory projection pursuit, maximum likelihood hebbian learning, self-organising maps, multiple classifier systems and Hybrid Artificial Systems and its applications to industrial and environmental problems.

## Lokaliais modeliais grįsta hibridinė regresijos sistema saulės energijos prognozavimui

Héctor QUINTIÁN, Jose Luis CALVO-ROLLE, Emilio CORCHADO

Šiame straipsnyje pateikto tyrimo tikslas yra energijos, kurią generuoja saulės šilumos sistemos, prognozavimas. Prognozavimui buvo sukurta hibridinė išmani sistema, grįsta lokalios regresijos nesudėtingais, didelio tikslumo modeliais. Pradiniai duomenys skirstomi į grupes, naudojant saviorganizuojančius neuroninius tinklus, o vėliau, kiekvienai grupei sukuriamas lokalus modelis. Atlikti tyrimai su įvairiais regresijos metodais. Nauja, lokaliais modeliais grįsta, hibridinė regresijos sistema buvo empiriniu būdu patikrinta, tyrimams naudojant bioklimato pastatų saulės šilumos sistemos duomenis.