Novel Approach for Estimating Monthly Sunshine Duration Using Artificial Neural Networks: A Case Study

Maamar Laidi1, Salah Hanini2, Abdallah El Hadj Abdallah3

1Laboratory of Biomaterials and Transport Phenomena (LBMPT), University of Médéa, BD de L’A.L.N Ain D’heb Médéa, Médéa, Algeria
e-mail: maamarw@yahoo.fr
2Laboratory of Biomaterials and Transport Phenomena (LBMPT), University of Médéa, BD de L’A.L.N Ain D’heb Médéa, Médéa, Algeria
e-mail: s_hanini2002@yahoo.fr
3Laboratory of Biomaterials and Transport Phenomena (LBMPT), University of Médéa, BD de L’A.L.N Ain D’heb Médéa, Médéa, Algeria
Department of Chemistry, University of Saad Dahlab, Route de Soumaa BP 270, Blida, Algeria
e-mail: a_abdallahelhadj@univ-blida.dz


ABSTRACT

This work deals with the potential application of artificial neural networks to model sunshine duration in three cities in Algeria using ten input parameters. These latter are: year and month, longitude, latitude and altitude of the site, minimum, mean and maximum air temperature, wind speed and relative humidity. They were selected according to their availability in meteorological stations and based on the fact that they are considered as the most used parameters by researchers to model sunshine duration using artificial neural networks. Several network architectures were tested to choose the most accurate and simple scheme. The optimum number of layers and neurons was determined by trial and error method. The optimized network was obtained using Levenberg-Marquardt back-propagation algorithm, one hidden layer including 25 neurons with Tan-sigmoid transfer function. The model developed in this study has the ability to estimate sunshine duration with a mean absolute percentage error value equals to 2.015%, a percentage root mean square error of 2.741% and a determination coefficient of 0.9993 during test stage.

KEYWORDS

Sunshine duration, Solar energy, Artificial neural networks, Root mean square error, Meteorological parameters.

INTRODUCTION

Sunshine duration is the period of time during which the ground surface receives radiation by direct sunlight [1]. This sunshine duration period is defined by the World Meteorological Organization (WMO) in 2003 to be the length of time during which direct solar irradiance exceeds a threshold of 120 W/m² [2].

* Corresponding author
According to Stambouli et al. [3], the sunshine duration over the whole Algerian territory exceeds 3,000 h annually and may reach 3,900 h in high plains and Sahara region. This huge quantity of sunshine per year makes Algeria one of the countries with the highest solar radiation levels in the world.

Knowledge of sunshine duration is mandatory for the design and optimization of renewable solar energy systems, but its measurements are always not available due to high costs and technical complexities.

Modeling and simulation of the nonlinear continuous variation of sunshine duration is to know how this parameter will behave without doing experimental measurements. Due to complexity of the relationship between sunshine duration and its dependent operating parameters, computational intelligence models are often more flexible against statistical models. Nowadays, a powerful method based on artificial intelligence namely Artificial Neural Networks (ANN), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Support Vector Regression (SVR) and Adaptive Neuro-fuzzy Inference System (ANFIS) in modeling complex phenomena has been successfully used [4, 5].

Therefore, various approaches have been proposed in literature to predict sunshine duration data in some parts of the world.

Mohandes and Rehman [1] made a comparison between two algorithms, namely, Particle Swarm Optimization (PSO) and Support Vector Machine (SVM) to estimate the sunshine duration of any region in Saudi Arabia. The best model was found with PSO algorithm.

Kabaa et al. [6] evaluated the potential of SVM approach for estimating daily sunshine duration using three different kernels of SVM, linear, polynomial, and Radial Basis Function (RBF). Results have shown that the SVM methodology can be a good alternative to those conventional and ANN methods for estimating daily sunshine duration.

Rahimikhoob [7] designed a Multi-Layer Perceptron (MLP) to estimate sunshine duration based on air temperature and humidity data under arid environment south-east of Iran. Results show that modeling by ANN technique made satisfactory estimates.

Kandirmaz et al. [8] uses an ANN approach for estimating monthly mean daily values of global sunshine duration for Turkey. Statistical indicators have shown that, Generalized Regression Neural Network (GRNN) and MLP models produced better results than the RBF model and can be used safely for the estimation of monthly mean daily values of global sunshine duration.

In comparison to aforementioned methods, ANN can be used to model many complex processes due to its robustness, simplicity, non-linearity and its ability to learn from experimental samples without any prior assumptions about their nature.

To this date, there is no comprehensive ANN model that is able to estimate sunshine duration. The novelty of this study is to investigate the feasibility of using three layers feed forward back propagation ANN algorithm to model the nonlinear continuous variation of monthly mean daily values of global sunshine duration in Algeria. These sunshine duration data are desirable for calculating many important parameters such as reference evapotranspiration and daily solar radiation.

**DESIGN OF THE NEURAL MODEL**

The ANN implementation consists of several stages which are explained in detail below and summarized in the flow chart shown in Figure 1 [9, 10]:

- In this study, data were arbitrarily split into 3 groups:
  - 33% for training stage (11 × 124);
  - 33% for testing stage (11 × 124);
  - 33% for validation stage (11 × 124);
• The training is carried out through the exposition of the network to a specific data set of information as well as by applying a training algorithm to produce the desired output from the network [11]. The efficiency of the trained network was tested and validated by using new sets of input not used in the training stage;

• Back-propagation networks refer to a multi-layered, feed-forward perceptron trained with an error back-propagation algorithm. Inputs are propagated forward through each layer to appear as outputs. The error function (i.e. objective function) between the latter and the desired ones are then propagated backward, then the weights and thresholds are adjusted to minimize the sum Square Error (MSE). Details of this algorithm are presented in the literature [12-14];

• In this study, Levenberg-Marquardt algorithm [15] was adopted to minimize the objective function;

• Despite the fact that there are some published methods showing how to initiate weights [16], which are important factors of ANN, these suggestions are not generalized;

• To improve the learning speed, a linear normalization in the range of $[-1, +1]$ has been adopted in this study using eq. (1) and eq. (2) hereinafter summarized in Table 1 [17];

• As ANNs are sensitive to the number of hidden neurons, their selection has to be done by trial and error method independently of the number of inputs [18]. The ANNs need to remain the simplest possible (parsimony principle) [19];

• Different types of transfer functions are proposed in the literature [15]. The sigmoid activation function [Table 1, eq. (3)] in the hidden layers is commonly used to introduce nonlinearity into the network and the purelin transfer function [Table 1, eq. (4)] can satisfy the requirement in the output layer [20, 21];

• Each topology was repeated several times to avoid the overfitting of the designed model caused by the initial random initialization of weights and thresholds [22, 23];

• Percentage Root Mean Square Error (%RMSE), Mean Absolute Percentage Error (MAPE) and Coefficient of determination ($R^2$), eq. (5) to eq. (7) respectively, illustrating the performance of the ANN model [24], are listed in Table 1;

• To optimize ANN parameters, all the aforementioned steps will be programmed using Matlab software [25].

Figure 1. Flow chart of ANN training processes [28]
Table 1. Summary of some mathematical equations used to calculate the ANN accuracy

<table>
<thead>
<tr>
<th>Eq.</th>
<th>Measure and Matlab code</th>
<th>Calculation</th>
<th>Nomenclature</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>Data normalizing equation (premnmx/tramnmx) [17, 26]</td>
<td>( Y = \frac{X - X_{\min}}{X_{\max} - X_{\min}} (Y_{\max} - Y_{\min}) )</td>
<td>( X ) is the original data value (input) and ( Y ) is the corresponding normalized variable. Tansig transfer function ranges between ([-1, +1]), this implies that ( Y_{\max} = -1 ) and ( Y_{\min} = 1 )</td>
</tr>
<tr>
<td>(2)</td>
<td>De-normalizing (to real value) of data estimated by ANN (postmnmx) [17, 26]</td>
<td>( X = \frac{Y (X_{\max} - X_{\min}) - Y_{\min}}{Y_{\max} - Y_{\min}} )</td>
<td>( X ) is the original data value (input) and ( Y ) is the corresponding normalized variable. Tansig transfer function ranges between ([-1, +1]), this implies that ( Y_{\max} = -1 ) and ( Y_{\min} = 1 )</td>
</tr>
<tr>
<td>(3)</td>
<td>Hyperbolic tangent sigmoid transfer function (tansig) [26]</td>
<td>( f(X) = \frac{2}{1 + e^{-2X}} - 1 = \frac{1 - e^{-2X}}{1 + e^{-2X}} = \frac{e^X - e^{-X}}{e^X + e^{-X}} )</td>
<td>( f ) denotes tansig transfer function</td>
</tr>
<tr>
<td>(4)</td>
<td>Linear function (purelin) [27]</td>
<td>( g(X) = X )</td>
<td>( g ) denotes purelin transfer function</td>
</tr>
<tr>
<td>(5)</td>
<td>Percentage Root Mean Square Error [8]</td>
<td>( %RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (T_i - P_i)^2} \times \frac{100}{\sum_{i=1}^{N} T_i} )</td>
<td>( T_i ) and ( P_i ) denote target and ANN predicted values respectively and ( N ) denotes the number of data points in the used set</td>
</tr>
<tr>
<td>(6)</td>
<td>Mean Absolute Percentage Error [1]</td>
<td>( MAPE(%) = \frac{100 \sum_{i=1}^{N}</td>
<td>T_i - P_i</td>
</tr>
<tr>
<td>(7)</td>
<td>Coefficient of determination [27]</td>
<td>( R^2 = \frac{\sum_{i=1}^{N} (T_i - \bar{T})^2 - \sum_{i=1}^{N} (T_i - P_i)^2}{\sum_{i=1}^{N} (T_i - \bar{T})^2} )</td>
<td>( \bar{T} ) is the average value of the experimental data</td>
</tr>
</tbody>
</table>

**SUNSHINE PREDICTION USING NEURAL NETWORK**

The neural network toolbox of Matlab software (Matlab Version 7.9.0 R2009b) has been employed to develop the desired feed-forward MLP network.

Due to paucity of data, ten parameters were selected as inputs for the ANN to model the sunshine duration in three sites in Algeria.

In this work, meteorological data were obtained from the National Office of Meteorology (ONM) Algiers-Algeria. Detailed geographic, climatic characteristics and source of data used for training/validation and testing of ANN model are given in Table 2. The experimental data gathered for the three selected cities were analyzed to delete outliers manually. This stage has been done by plotting the variation of each parameter in Matlab software. Results revealed that about 2% of data lie outside the expected range due to the equipment failure and other technical problems.

<table>
<thead>
<tr>
<th>Sites</th>
<th>Latitude (N°)</th>
<th>Longitude (E°)</th>
<th>Altitude [m]</th>
<th>Climate [3, 29]</th>
<th>Points</th>
<th>Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Médéa</td>
<td>36.45</td>
<td>2.75</td>
<td>1,030</td>
<td>Mediterranean</td>
<td>132</td>
<td>2000-2009 and 2011</td>
</tr>
<tr>
<td>Tlemcen</td>
<td>35.01</td>
<td>1.28</td>
<td>246</td>
<td>Mediterranean</td>
<td>120</td>
<td>1996-2005</td>
</tr>
<tr>
<td>Batna</td>
<td>35.45</td>
<td>6.19</td>
<td>822</td>
<td>Semi-arid climate</td>
<td>120</td>
<td>1996-2005</td>
</tr>
</tbody>
</table>

The distribution of measured sunshine duration data has been depicted in Figure 2.
RESULTS AND DISCUSSION

The feed forward ANN type with Levenberg-Marquardt Back Propagation training algorithm was chosen because it is suitable for modeling the relationship between input and output variables [28].

In this paper, the performance of ANN with varying number of neurons (1-35) in the hidden layer was investigated where each ANN-architecture was trained 10 times in order to avoid random effects. Tests have been performed on two hidden layers and with different parameters. However, no improvements have been reported.

Results demonstrated that architectures with one hidden layer were able to reach the goal in term of errors and determination coefficient. Table 3 shows the structure of the optimized ANN model. The accuracy of the networks was evaluated for each epoch in the training through Mean Squared Error (MSE). The best validation performance is \(10^{-3}\) at epoch (iterations) 1,000 for the best network topology with overall MAPE of 2.0152\% and \%RMSE of 2.7411\%, indicating an accurate mapping of the data.

Table 3. Architecture of the optimized ANN model

<table>
<thead>
<tr>
<th>Type of network</th>
<th>Training algorithm</th>
<th>Input layer</th>
<th>Hidden layer number: 01</th>
<th>Output layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFBP NN (newff Matlab function)</td>
<td>Levenberg-Marquardt backpropagation (trainlm Matlab function)</td>
<td>10</td>
<td>25</td>
<td>Hyperbolic tangent sigmoid (tansig Matlab function)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Linear (purelin Matlab function)</td>
</tr>
</tbody>
</table>

The performance of the designed ANN is given in Table 4.

Table 4. Performances of MLP

<table>
<thead>
<tr>
<th></th>
<th>MAPE [%]</th>
<th>%RMSE [%]</th>
<th>(R^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>1.9547</td>
<td>2.6236</td>
<td>0.9994</td>
</tr>
<tr>
<td>Test</td>
<td>2.2567</td>
<td>2.9374</td>
<td>0.9991</td>
</tr>
<tr>
<td>Validation</td>
<td>1.8343</td>
<td>2.4557</td>
<td>0.9994</td>
</tr>
<tr>
<td>All</td>
<td>2.0152</td>
<td>2.7411</td>
<td>0.9993</td>
</tr>
</tbody>
</table>

Figure 3 shows a comparison between experimental and predicted values of sunshine duration during training and test stage. As this plot shows, there is very good agreement between target data and the results obtained from the ANN model with an overall \(R^2\) of 0.9993. The \(R^2\) value for the sunshine duration is close to 1.00 for the training, validation, test and overall data sets showing that the model has captured the features quite accurately. Results revealed that an accurate correlation exists between the correlated data by ANN model and experimental ones.

Figure 4 shows a comparison between experimental data and NN predicted results. This figure shows an excellent agreement between real data (shown as white face markers) and the NN predicted results (shown as colored face markers). The predicted data curves follow the same trend of the experimental data with slight difference. Overall, results show a good predictive ability of the NN model for sunshine duration estimation. The proposed approach can be used as a method to estimate sunshine duration in the remote and rural locations in Algeria with no direct measurement devices at least in
An accurate preliminary statistical analysis of the data used in this study with maximums, minimums, standard deviations, standard error, variances, kurtosis and skewness has been done in Table 5.

Figure 3. Experimental vs. predicted sunshine duration: for the training data set (a); for the test data set (b); for the validation data set (c) and overall data set (d)
Figure 4. Comparison between experimental data and NN predicted results during the test stage: Batna city (a); Tlemcen (b) and Médéa city (c)

Table 5. Input and output parameters used in modeling

<table>
<thead>
<tr>
<th>Variable category</th>
<th>No.</th>
<th>Parameters</th>
<th>Unit</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Standard deviations</th>
<th>Standard error</th>
<th>Variances</th>
<th>Kurtosis</th>
<th>Skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input</strong></td>
<td>1</td>
<td>Year</td>
<td>[-]</td>
<td>1996</td>
<td>2011</td>
<td>0.2</td>
<td>0.2</td>
<td>14.1</td>
<td>2.564</td>
<td>0.322</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Month number</td>
<td>[-]</td>
<td>1</td>
<td>12</td>
<td>0.2</td>
<td>0.2</td>
<td>11.9</td>
<td>1.783</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Longitude</td>
<td>['°']</td>
<td>35.01</td>
<td>36.45</td>
<td>0.6</td>
<td>0.6</td>
<td>0.4</td>
<td>1.413</td>
<td>0.353</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Latitude</td>
<td>['°']</td>
<td>1.28</td>
<td>6.19</td>
<td>0.6</td>
<td>0.6</td>
<td>0.4</td>
<td>1.413</td>
<td>0.353</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Altitude</td>
<td>[m]</td>
<td>246</td>
<td>1,030</td>
<td>331.9</td>
<td>17.2</td>
<td>110.128</td>
<td>1.761</td>
<td>−0.553</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>Minimum air temperature</td>
<td>[°C]</td>
<td>−3.6</td>
<td>22.3</td>
<td>6.6</td>
<td>0.6</td>
<td>43.9</td>
<td>1.801</td>
<td>0.174</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>Average air temperature</td>
<td>[°C]</td>
<td>2.9</td>
<td>29.2</td>
<td>7.4</td>
<td>0.4</td>
<td>55.0</td>
<td>1.653</td>
<td>0.189</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>Maximum air temperature</td>
<td>[°C]</td>
<td>5.7</td>
<td>39.3</td>
<td>8.7</td>
<td>0.5</td>
<td>76.1</td>
<td>1.761</td>
<td>0.094</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>Wind speed</td>
<td>[m/s]</td>
<td>0.8</td>
<td>5.7</td>
<td>0.9</td>
<td>0.0</td>
<td>0.8</td>
<td>2.840</td>
<td>0.497</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>Relative humidity</td>
<td>[%]</td>
<td>32</td>
<td>89</td>
<td>12.7</td>
<td>0.7</td>
<td>161.3</td>
<td>2.593</td>
<td>−0.502</td>
</tr>
<tr>
<td><strong>Output</strong></td>
<td>11</td>
<td>Sunshine duration</td>
<td>[h]</td>
<td>91</td>
<td>390.2</td>
<td>66.1</td>
<td>3.4</td>
<td>4,370.6</td>
<td>2.155</td>
<td>−0.059</td>
</tr>
</tbody>
</table>

The developed model is given by the following formula [eq. (8)]:

\[
SD = \sum_{s=1}^{S} \left[ W_{0(l,s)} \left( \frac{2}{1 + \exp(-2\sum_{k=1}^{K} (W_{i(s,k)} \cdot \ln_{(k)} + b_{1(s)}) + b_{2(l)})} - 1 \right) \right] + b_{2(l)}
\]

(8)

where \( s \) is the number of neurons in the hidden layer (\( S = 25 \)), \( k \) is the number of neurons in the input layer (\( K = 10 \)), \( l \) is the number of neurons in output layer (\( l = 1 \)), \( W_i \), \( W_0 \) and \( b_{1(s)} \), \( b_{2(l)} \) are weights and biases, respectively. The major disadvantage of this technique is that it can be used only in the range in which it has been trained as it is empirical in nature. For this reason, ranges of the used database for the model are shown in Table 5.

Once a satisfactory degree of input-output mapping was achieved, the network training was stopped and a set of completely unknown test data was applied for verification.

The performance of the herein developed ANN model is compared to the previously developed ones in literature for prediction of sunshine duration [1, 7, 8]. Details of the
comparison are given in Table 6 in terms of $R^2$ and different errors. Based on the obtained results, the determination coefficient and the different errors obtained in the present study are very acceptable than those ones published in literature.

Table 6. Review of the proposed ANN models for sunshine duration estimation using ANN

<table>
<thead>
<tr>
<th>Authors</th>
<th>Type</th>
<th>$R^2$ test</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>Two NN algorithms:</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PSO algorithm 0.612</td>
<td>2.3%-22.9% overall average of 6.9%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SVM 0.668</td>
<td>2.7%-16.7% overall average of 6.5%</td>
<td></td>
</tr>
<tr>
<td>[7]</td>
<td>MLP 0.81</td>
<td>%RMSE of 6.3%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GRNN 0.9530</td>
<td>9.88</td>
<td></td>
</tr>
<tr>
<td>[8]</td>
<td>MLP 0.9563</td>
<td>9.36</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RBF 0.8902</td>
<td>10.75</td>
<td></td>
</tr>
<tr>
<td>[6]</td>
<td>SVM-RBF 0.8435</td>
<td>$RMSE = 1.5105$ h</td>
<td></td>
</tr>
<tr>
<td>This study</td>
<td>MLP 0.9993</td>
<td>Overall average MAPE of 2.0125%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overall average %RMSE of 2.7411%</td>
<td></td>
</tr>
</tbody>
</table>

CONCLUSIONS

In this study, a multilayer feedforward neural network with a learning scheme of the BP of errors and the Levenberg-Marquardt algorithm for the adjustment of the connecting weights were applied for modeling ground measurement sunshine duration. The best network topology was obtained as (10-25-1) with the hyperbolic sigmoid as transfer function for the hidden layers and linear transfer function of the output layer. The neural network model presently developed gives an overall MAPE of 2.015% and %RMSE of 2.741% which led to a significantly better estimation of sunshine duration. This was further confirmed by the value of coefficient of determination which is 0.9993.

After the comparison between the predicted results and those of the four published models, it has been shown that the proposed ANN models have a better performance than the literature models. The optimized ANN can be used in these three cities to:

- Complete some missing data due to heavy power-cuts especially in summers;
- Complete data when it lies outside the expected range due to equipment failure and other problems like (calibration problems, dirt on the sensor, accumulated water, shading of the sensor by masts);
- Design and optimize solar appliances;
- Dispense with the use of a heliograph and finally.

As no model can be considered as the best one and with more available experimental data the model can be improved to be applicable for many cities in Algeria.

ACKNOWLEDGMENT

The authors would like to thank the National Office of Meteorology, Algiers-Algeria for providing data used in this study. Anonymous reviewers are gratefully acknowledged for their helpful suggestions that have led to improving the present manuscript.

REFERENCES


413


Paper submitted: 02.06.2017
Paper revised: 08.06.2018
Paper accepted: 09.06.2018