COMPARISON OF REGRESSION AND NEURAL NETWORKS MODELS TO ESTIMATE SOLAR RADIATION

Mónica Bocco¹*, Enrique Willington¹, and Mónica Arias²

ABSTRACT

The incident solar radiation on soil is an important variable used in agricultural applications; it is also relevant in hydrology, meteorology and soil physics, among others. To estimate this variable, empirical models have been developed using several parameters and, recently, prognostic and prediction models based on artificial intelligence techniques such as neural networks. The aim of this work was to develop linear models and neural networks, multilayer perceptron, to estimate daily global solar radiation and compare their efficiency in its application to a region of the Province of Salta, Argentina. Relative sunshine duration, maximum and minimum temperature, rainfall, binary rainfall and extraterrestrial solar radiation data for the period 1996-2002, were used. All data were supplied by Experimental Station Salta, Instituto Nacional de Tecnología Agropecuaria (INTA), Argentina. For both, neural networks models and linear regressions, three alternative combinations of meteorological parameters were considered. Good results with both prediction methods were obtained, with root mean square error (RMSE) values between 1.99 and 1.66 MJ m⁻² d⁻¹ for linear regressions and neural networks, and coefficients of correlation ($r^2$) between 0.88 and 0.92, respectively. Even though neural networks and linear regression models can be used to predict the daily global solar radiation appropriately, neural networks produced better estimates.

Key words: modeling, prediction, linear regression, multilayer perceptron.

INTRODUCTION

The incident solar radiation on soil is an important variable used in agricultural applications, particularly for modeling crop development, values of soil moisture, potential evapotranspiration and photosynthesis, among others. It is also important in hydrology, meteorology and soil physics. Moreover, the availability of these data, or their estimation based on specific sites or mechanistic prediction models, improves the usefulness of the climate data sets (Ball et al., 2004).

In places where radiation measurements are sparse, theoretical estimations of the available solar energy can be used to predict these measurements from standard weather parameters that are extensively measured (air temperature, relative humidity, effective sunshine duration and cloudiness) (Santamouris et al., 1999).

While solar energy data are recognized as very important, their acquisition is not easy. The measurement of solar radiation requires the use of expensive equipment, and in developing countries there are not always adequate facilities to mount viable monitoring programs. Therefore, there have been several attempts to estimate solar radiation through the use of meteorological and physical parameters (Togrul and Togrul, 2002). The lack of observed atmospheric variables prevents the use of many analytical procedures and forces us to use their estimation by different methods in order to be able to use these procedures (De la Casa et al., 2003).

Several empirical models have been developed to calculate global solar radiation using various parameters, the relative sunshine duration is the most commonly used. In 1924 Angstrom used a linear relationship between global radiation and sunshine duration; a modified version of this correlation, proposed by Prescott in 1940, has been the most convenient and widely used for estimating global solar radiation; this is known as the Angstrom-Prescott equation (Podestá et al., 2004).

Almorox et al. (2008) adequately estimated global solar radiation for 11 meteorological stations in Venezuela from sunlight data using a linear regression to calculate
the Angstrom-Prescott equation. Falayi et al. (2008) developed, for Nigeria, multilinear regression equations to predict the relationship between global solar radiation with different weather parameters.

A simple and fast physically based method for the estimation of global solar radiation using meteorological satellite data for was presented for Wloczkyk and Richter (2006). For irrigated agricultural area was analyzed the distribution of net radiation flux density using a method that combine satellite remote sensing with field observation (Folhes et al., 2006).

Most of the studies used to predict solar radiation were based on time series methods (including regression analysis), which are limited in the number of parameters that can accurately handle. In particular, Fortin et al. (2008) developed a long-utilized linear approach based on latitude and daily temperature range. In addition, estimations of daily radiation resulting from an Angstrom-Prescott relationship have adequate accuracy at a monthly scale, but are not accurate at a daily scale (Ceballos et al., 2005).

Recently, prognostic and prediction models based on artificial intelligence techniques such as neural networks (NN) have been developed. These models can handle a large number of data, predict the contribution of these in the outcome and provide prompt and adequate predictions (Al-Alawi and Al-Hinai, 1998). Using neural networks, Bocco et al. (2006) made models to estimate solar radiation at Córdoba (Argentina), Mohandes et al. (1998) for Saudi Arabia and Fortin et al. (2008) for Canada.

Within this methodology, the multilayer perceptron is probably the most commonly used algorithm with the architecture of neural networks because of its capacity to tolerate information that is incomplete, inaccurate or contaminated with noise (Mas and Flores, 2008). The multilayer perceptron consists of a non-parametric statistical model of nonlinear regression which generally uses a single hidden layer to completely divide the spectral space by means of hyperplanes along which the level of activation of hidden units is constant (Foody, 2000).

The aim of this work was to develop linear models and neural networks to estimate daily global solar radiation from commonly observed meteorological data and compare the overall efficiency of these models and networks in an application to a region of the Province of Salta (Argentina).

MATERIALS AND METHODS

Site of application
Daily values of meteorological variables, including radiation, for the 1996-2002 period, were provided by the Experimental Station Salta (24°54’ S, 65°29’ W, 1234 m.a.s.l.), Instituto Nacional de Tecnología Agropecuaria (INTA), Argentina. All data were collected with an automatic weather station, Vantage Pro2 Stations (Davis Instruments, Hayward, California, USA). The agro meteorological station is part of the National Climate Network and takes weather observations three times a day. As regards to the type and location of the instruments with which samples are taken, both are standardized by the World Weather Organization and the National Meteorological Service (Estación Experimental Agropecuaria Salta, INTA). The astronomical solar radiation corresponding to this site was calculated using the Solar-Calc software by USDA-ARS (2007).

Linear models
The statistical analysis began studying the observed radiation distribution. There were 2550 observations, with an average value equal to 14.19 MJ m⁻² d⁻¹, minimum and maximum values equal to 1.20 and 28.80 MJ m⁻² d⁻¹, respectively. A coefficient of asymmetry with value -0.07 and percentiles 25 and 75 for this variable were equal to 10.29 and 18.50 MJ m⁻² d⁻¹, respectively.

For the variable under study, there were extreme values (minimum and maximum) and concentration of the values close to the average.

The meteorological parameters used to estimate solar radiation from the statistical correlations were maximum and minimum temperatures (°C), rainfall (mm), binary rainfall (a binary function with value 1 for occurrence and 0 for days with no precipitation), relative sunshine duration (RSD) and astronomical solar radiation (MJ m⁻² d⁻¹). For all variables we performed a correlation analysis to obtain a measure of the magnitude and direction of the association of each pair of variables. Since in Argentina, many stations only have instruments to measure and record some meteorological variables; it is a very useful tool to consider rainfall a binary variable.

For linear regression analysis three possible parameter combinations were considered: Regression R1: daily values of maximum temperature (Tmax), minimum temperature (Tmin), rainfall (R), relative sunshine duration (RSD) and astronomical solar radiation (ASR); Regression R2: daily values of maximum temperature, minimum temperature, binary rainfall (BinR), relative sunshine duration and astronomical solar radiation; and Regression R3: daily values of maximum temperature, minimum temperature, rainfall and astronomical solar radiation.

Neural networks models
A neural network (NN) model, multilayer perceptron, was used to estimate the incident solar radiation. This procedure is a mathematical model that performs a
computational simulation of the behaviour of neurons in the human brain by replicating, on a small scale, the brain’s patterns in order to produce results from the events perceived, i.e. it is a model based on learning a set of training data. The main characteristic of NN is their capacity for learning by example. This means that by using a NN there is no need to program how the output is obtained, given certain input; the NN will learn the existing input-output relationship by means of a learning algorithm. This learning will materialize in the network’s topology and in the value of its connections. Once the NN has learnt to carry out the desired function, input values for which the output is unknown can be entered, and the NN will calculate the output.

The NN are composed of a number of interconnected processing elements which are joined by weighted connections. The training algorithm adjusts the connection weights through an iterative procedure in which the error is minimized (Ashish et al., 2004). The amount of training data required for successful classification increases exponentially with increased dimensionality of the input data (Dixon and Candade, 2008). The Multilayer Perceptron (Figure 1) is a fully connected multilayer feed forward supervised learning network with symmetric hyperbolic tangent activation functions, trained by the back-propagation algorithm to minimize a quadratic error. The general steps that describe the training algorithm of the proposed networks are described, according to Bocco et al. (2006), as follows: Initialize the weights in the net with random values (step 1); read an input pattern \( X_p: (x_{p1}, x_{p2}, ..., x_{pN}) \) and the desired output \( d \) (step 2); generate the output calculated by the net for the presented input. To do so, the values of the answers in each layer are obtained, until the output layer is reached (step 3). The net for the hidden neurons (\( H_j \)) coming from the input (\( net \)) is calculated as follows:

\[
net_{pj}^h = \sum_{i=0}^{N} w_{ji}^h x_{pi} + \theta_{j}^h \tag{1}
\]

where the sub-index \( p \) corresponds to the \( p \)-th training vector, \( j \) to the \( j \)-th hidden neuron, \( w_{ji} \) is the weight of the connection between \( I_i \) and \( H_j \) and the term \( \theta_{j} \) corresponds to a term of the minimum threshold to be achieved by the neuron for its activation. Based on these inputs the outputs of the hidden neurons are calculated, using an activation function \( f \):

\[
y_{pj}^h = f_j^h (net_{pj}^h) \tag{2}
\]

To obtain the results of neuron in the output layer, the same is done:

\[
net_{pk}^o = \sum_{j=1}^{L} w_{kj}^o y_{pj}^h + \theta_{k}^o \tag{3}
\]

\[
y = f_k^o (net_{pk}^o) \tag{4}
\]

Once all neurons have an activation value for a given input pattern, the algorithm continues calculating the error for each neuron, except for those in the input layer (step 4). For the neuron in the output layer, if the answer is \( y \), such error (\( \delta \)) can be expressed as:

\[
\delta_{pk} = (d - y) f_k^o (net_{pk}^o) \tag{5}
\]

If the neuron \( j \) is not an output one, then the derivative of the error cannot be directly calculated. The error in the hidden layers depends on all the terms of the error in the output layer. For this reason they are called backpropagation.

In order to update the weights the recursive algorithm, starts with the output neuron and working backwards until the input layer is reached (step 5). This process is repeated an \( n \) number of times, so that an acceptably low square error (\( E_p \)) for all the learned patterns, can be reached (step 6).

Figure 1. Schematic map of a multilayer perceptron artificial neural network.
In our work, the size of the input layer that receives the information from various parameters that affect the radiation is the number of variables (described in detail later) and the output layer has one neuron which indicates the predicted total daily solar radiation (Est Rad). The number of neurons in the hidden layer and the number of hidden layers are selected during the training process.

The final process of this technique is the validation that always requires a separate data set for which we know the phenomenon behaviour and on which errors are estimated. The aim was to verify the efficiency of the designed NN.

The training process used 50% of the data, these were taken at random from the 1996-2002 period and 2000 iterations were performed. The validation process was carried out with the other half of the data, all corresponding to Salta. To evaluate the models’ performance, the statistical parameters root mean squared error (RMSE) and correlation coefficient ($r^2$), were considered.

The use of neural networks (NN) has opened new perspectives since they do not hypothesize on data distribution (Walthall et al., 2004). It was verified that the observed solar radiation does not correspond to a normal distribution using a Shapiro-Wilks test (modified) and a Kolmogorov test for goodness-of-fit ($p < 0.05$ in both tests).

NN models considered three alternatives of the variables for the input layer, equivalent to the parameter combinations in linear regressions: (1) Model M1: daily values of Tmax, Tmin, rainfall, ASR and ASR (the same parameters as R1); (2) Model M2: daily values of Tmax, Tmin, binary rainfall, RSD and ASR (analogous to R2); and (3) Model M3: daily values of Tmax, Tmin, rainfall and ASR (parameters of R3).

The three models were constructed with an input layer of four (M3) or five (M1 and M2) neurons and one hidden layer of 10 neurons.

With the aim of comparing the linear regression results with the developed NN model results, correlations and regressions were performed using only half the data, exactly the same data set used in the training phase of the NN.

### RESULTS AND DISCUSSION

The results of the validation process of all models allowed the calculation of different statistic values between observed and estimated values of solar radiation (Table 1).

When studying the results of the correlation analysis, there was a correlation between the observed radiation (Ob Rad) and the variables ($p$-values < 0.05 in row 1 Table 2), except for rainfall. The coefficients between the observed radiation and the dependent variables analyzed (column 1), point to a positive correlation of various sizes for Tmax, RSD and ASR, and negative value for binary rainfall (Table 2).

In the scatter plots of Figure 2 the relationship between the observed radiation and the other variables is shown. The graphics displayed a high correlation with the RSD, a low correlation with Tmin, no correlation with rainfall and a high correlation with the Tmax,

### Table 1. Root mean squared error (RMSE) and correlation coefficient ($r^2$) in the validation phase for the different models.

<table>
<thead>
<tr>
<th>Models</th>
<th>RMSE</th>
<th>$r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>1.66</td>
<td>0.92</td>
</tr>
<tr>
<td>M2</td>
<td>1.68</td>
<td>0.91</td>
</tr>
<tr>
<td>M3</td>
<td>2.97</td>
<td>0.73</td>
</tr>
<tr>
<td>R1</td>
<td>2.02</td>
<td>0.88</td>
</tr>
<tr>
<td>R2</td>
<td>1.99</td>
<td>0.88</td>
</tr>
<tr>
<td>R3</td>
<td>3.44</td>
<td>0.64</td>
</tr>
</tbody>
</table>

### Table 2. Correlation matrix between solar radiation and other meteorological variables.

<table>
<thead>
<tr>
<th>Ob Rad</th>
<th>Tmax</th>
<th>Tmin</th>
<th>R</th>
<th>Bin R</th>
<th>RSD</th>
<th>ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ob Rad</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.76</td>
<td>1.1E-12</td>
<td>0.00</td>
</tr>
<tr>
<td>Tmax</td>
<td>0.72</td>
<td>1.00</td>
<td>0.00</td>
<td>1.7E-6</td>
<td>0.50</td>
<td>0.00</td>
</tr>
<tr>
<td>Tmin</td>
<td>0.25</td>
<td>0.52</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>R</td>
<td>-0.01</td>
<td>0.14</td>
<td>0.23</td>
<td>1.00</td>
<td>0.00</td>
<td>2.6E-4</td>
</tr>
<tr>
<td>Bin R</td>
<td>-0.21</td>
<td>-0.02</td>
<td>0.34</td>
<td>0.44</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>RSD</td>
<td>0.76</td>
<td>0.51</td>
<td>-0.23</td>
<td>-0.11</td>
<td>-0.39</td>
<td>1.00</td>
</tr>
<tr>
<td>ASR</td>
<td>0.51</td>
<td>0.54</td>
<td>0.73</td>
<td>0.21</td>
<td>0.29</td>
<td>-0.05</td>
</tr>
</tbody>
</table>

Ob Rad: observed radiation; Tmax: maximum temperature; Tmin: minimum temperature; R: rainfall; BinR: binary rainfall; RSD: relative sunshine duration; ASR: astronomical solar radiation.
although this correlation does not correspond to a linear model.

The regression coefficients for linear models R1, R2 and R3 were:

\[
R_1 = -6.22 + 0.03 T_{\text{max}} + 0.04 T_{\text{min}} - 0.04 R + 0.15 \text{ RSD} + 0.5 \text{ ASR}
\]

\[
R_2 = -5.97 + 0.02 T_{\text{max}} + 0.05 T_{\text{min}} - 1.18 \text{ Bin } R + 0.15 \text{ RSD} + 0.51 \text{ ASR}
\]

\[
R_3 = -8.28 + 0.7 T_{\text{max}} - 0.47 T_{\text{min}} - 0.08 R + 0.46 \text{ ASR}
\]

These regression correlation values ranged between \( r^2 = 0.88 \) and 0.64. For Nigeria, Falayi et al. (2008) found, when they related the ratio between observed and astronomical radiation, correlation values ranging between \( r^2 = 0.56 \) and 0.97 according to the construction of regression equations with respect to a single variable or a combination between RSD, ratio of minimum and maximum temperature, relative humidity and monthly average daily temperature.

The NN models presented high correlation values, in particular M1 obtained a RMSE = 1.66 MJ m\(^{-2}\) d\(^{-1}\) and M2 got a RMSE = 1.68 MJ m\(^{-2}\) d\(^{-1}\) and RMSE = 2.97 MJ m\(^{-2}\) d\(^{-1}\) for M3; consequently, M1 and M2 can be used to make good estimates of daily global solar radiation values from registered data of daily maximum and minimum temperature, rainfall (or binary rainfall for M2), RSD and theoretical ASR.

In order to analyze the performance of the models that present better adjustment (M1, M2, R1 and R2), scatter plots considering observed and estimated solar radiation values were done (Figure 3).

The obtained results are considered a good estimate of global solar radiation because they are consistent with those published by other authors. Podestá et al. (2004) for the Humid Pampa, applying the Angstrom-Prescott equation, reported RMSE between 1.54 and 1.90 MJ m\(^{-2}\) d\(^{-1}\) using RSD, and when they used temperature and precipitation RMSE increased to 3.23 and 4.28 MJ m\(^{-2}\) d\(^{-1}\).

For Canada, Fortin et al. (2008) developed a multiple-layer perceptron network (same kind to the NN used in this work) to estimate surface incoming solar radiation on an horizontal surface, obtaining, with different input variables, RMSE between 3.83
and 5.45 MJ m\(^{-2}\). Using NN for Cordoba (Argentina), Bocco et al. (2006), with thermal amplitude, rainfall, cloudiness and RSD data, obtained RMSE similar to those estimated for Salta, with values ranging between 3.15 and 3.88 MJ m\(^{-2}\) d\(^{-1}\).

In the analyzed models, the temporal evolution of the calculated radiation values shows a seasonal pattern that fits correctly to annual variation of solar radiation. As an example, Figure 4 shows the temporal evolution of the values estimated by model M1.

The results show that M1 and M2, for NN models, and R1 and R2 for linear regression, have the lowest RMSE values. These have also the highest correlation coefficients (Table 1). Comparing the statistics of M1, M2 and M3 with R1, R2 and R3, respectively, smaller values of error and higher correlation coefficients for neural networks were observed. Surely this could be due to the nonlinearity of the relationship of solar radiation with any of the considered variables, and as noted by Verger et al. (2008) NN allow good estimates for complex and nonlinear problems.

The RMSE and \(r^2\) values of both M3 model and R3 show the importance of RSD data to estimate the total daily solar radiation, because although the coefficient \(r^2 = 0.73\) for M3 indicates a proper estimation without this information, better results are obtained when this parameter is included in the models, a similar behaviour is observed on linear regressions.

**CONCLUSIONS**

Solar radiation can be adequately estimated by linear models and neural networks, from values of meteorological variables of routine use; even NN produced better estimates.

Neural networks are an efficient methodology to estimate daily solar radiation, using a reduced number of meteorological parameters; they allowed, principally, reproduce the solar radiation evolution patterns for Salta (Argentina).

Even though linear regressions produce good estimates of daily global solar radiation, predictions are strongly correlated to the data set used.

Relative sunshine duration is a key variable involved in the calculation procedures of several agricultural and environmental indices. Estimation of surface incoming
solar radiation is, therefore essential, and models such as the one proposed might prove extremely useful.

ACKNOWLEDGEMENT

The production of this manuscript was supported in part by (Secretaría de Ciencia y Tecnología de la Universidad Nacional de Córdoba). The authors are grateful to meteorologist Ignacio Nieva (Estación Experimental Agropecuaria - INTA Cerrillos, Salta, Argentina) for providing data used in this paper.

RESUMEN

Comparación de regresión y modelos de redes neuronales para estimar la radiación solar. La radiación solar incidente en el suelo es una variable importante usada en aplicaciones agronómicas, además es relevante en hidrología, meteorología y física del suelo, entre otros. Para estimarla se han desarrollado modelos empíricos que utilizan distintos parámetros meteorológicos y, recientemente, modelos de pronóstico y predicción basados en técnicas de inteligencia artificial tales como redes neuronales. El objetivo de este trabajo fue desarrollar modelos lineales y de redes neuronales, del tipo perceptrón multicapa, para estimar la radiación solar global diaria y comparar la eficiencia de los mismos en su aplicación para una región de la Provincia de Salta, Argentina. Se utilizaron datos de heliofanía relativa, temperaturas máxima y mínima, precipitación, precipitación binaria y radiación solar astronómica provistos por la Estación Experimental Salta, Instituto Nacional de Tecnología Agropecuaria (INTA), Argentina, correspondientes al periodo 1996-2002. Tanto para los modelos de redes neuronales como para las regresiones lineales se consideraron tres alternativas de combinaciones de los parámetros meteorológicos, obteniéndose buenos resultados con ambas metodologías de predicción, con valores de la raíz del error cuadrático medio variando desde 1.99 a 1.66 MJ m⁻² d⁻¹ y coeficientes de correlación de 0.88 a 0.92. Se concluye que ambos, los modelos de redes neuronales y las regresiones lineales, pueden ser usados para predecir en forma adecuada la radiación solar global diaria; si bien las redes neuronales produjeron mejores resultados.

Palabras clave: modelos, predicción, regresiones lineales, perceptrón multicapa.

LITERATURE CITED


