International Environmental Modelling and Software Society (iEMSs)
2010 International Congress on Environmental Modelling and Software
Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada
David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.)
http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings

Integration of Statistical and Machine Learning Models for Short-term Forecasting of the Atmospheric Clearness Index

L. Mora-Lopez [a], **M. Piliougine**[b], **J.E. Carretero**[b] **and M. Sidrach-de-Cardona**[b]

[a]*Dpto. Lenguajes y C.Computación,ETSI Informática. Universidad de Málaga. Spain (llanos@lcc.uma.es)*

[b]*Dpto. Física Aplicada II. EU Politécnica. Universidad de Málaga. Spain (michel@ctima.uma.es, carretero@ctima.uma.es, msidrach@ctima.uma.es)*

Abstract: We have developed a framework that integrates statistical and machine learning models for the short-term forecasting of a climatic parameter know as the atmospheric clearness index. We have used a multivariate regression to establish the most significant variable amongst all the previous values for the clearness index series. The value of this variable was used to divide all the available data in several intervals. Three different models were used to select the model that best fits the observations for each of these intervals. With this set of models, one for each interval, we have built a framework that enables the short term forecasting of the clearness index. In this framework, the best model is used for each prediction taking into account the current value of this parameter. Data from 10 Spanish locations have been used to build the framework and to check the forecasting accuracy. The results show that prediction involving different models is better than when only one model is used.

Keywords: machine learning; atmospheric clearness index; forecasting.

1 INTRODUCTION

The atmospheric clearness index, also known as atmospheric transmissivity or atmospheric transmission coefficient, Kt is the ratio of global solar radiation at ground level to extra-terrestrial solar radiation:

$$K_t = \frac{G_t}{G_{0,t}} \tag{1}$$

where t is the period of time, for instance, hourly or daily.

The prediction of the clearness index is used as a step prior to estimating the global solar radiation availability that is the input data used for sizing and evaluating solar energy systems such as thermal or photovoltaic.

The prediction and simulation of the clearness index can be performed on different time scales: long, medium and short term. Long-term prediction allows the prediction of the mean and variance of real series to be obtained and the sequential properties observed in them to be reproduced. Medium-term prediction derives an estimated time distribution for the solar radiation day. Short-term prediction provides the prediction for the next interval time (daily, hourly,...).

Short-term prediction of global solar radiation -that is directly obtained from the clearness index value using the extraterrestrial solar radiation- is important for the following, among other, tasks:

- Management of grid-connected photovoltaic systems, particularly for medium and large ones.
- Management of medium and high temperature collectors of thermal plants, such as parabolic or power tower systems.
- Environmental control of greenhouses and solar driers: to estimate the heat energy that is expected to be available from the Sun and, therefore, to better compute the heating and ventilation needs in order to achieve minimization of the energy consumption.

In the short-term forecasting, expected clearness index values and, therefore, of global solar radiation will be highly useful for some systems that use solar energy as a resource. For instance, the accurate prediction of the energy produced by an photovoltaic or a power tower system (thermal energy) can help producers to achieve optimal management and can also help to use efficient operation strategies and to decide the best way of interacting with the conventional grid.

However, short-term prediction of the clearness index at earth surface is a complex problem as it depends on the atmospheric components affecting solar radiation as it passes through the atmosphere. The cloud presence is the most important of these factors in terms of the attenuation of solar radiation. The problem is that the cloud attenuation process is highly stochastic in nature and it is therefore difficult to predict how it will affect solar radiation. The attenuation process is nonlinear, complex, dynamic and widely scattered due to the influence of physical phenomena involved and to their variability in space and time.

Statistical time series models have been used for forecasting the atmospheric clearness index. Basically, the approach presented in Box and Jenkins [1976] is used. These models are particularly useful for long term characterization and prediction of the clearness index as they pick up the statistical and sequential properties of this parameter. Some of the most widely accepted methods use Autoregressive and Moving Average models to characterize and simulate the hourly and daily series of clearness index, Brinkworth [1977], Bartoli [1983], Aguiar [1988], Graham [1988], Aguiar and Collares-Pereira [1992], Mora-Lopez [1998]. However, these methods have not been used for short-term prediction of clearness index as the error in the prediction of isolated values (next value in a series) is too large.

In recent years, new forecasting models based in different machine learning approaches have been proposed to overcome the limitations of statistical methods. Machine learning models do not require any assumptions to be made as in the case of statistical methods, particularly with respect to the linearity of the series. Some of the recent machine learning models for different climatic parameters can be found in Elminir [2007] , Kilsby [2007], Al-Alawi [2008],Ingsrisawang [2008], Cao and Lin [2008], Mubiru [2008], Ghanbarzadeh [2009].

To address the need for short term forecasting of the clearness index, this paper proposes the use of models both from the statistical and machine learning areas. The paper is organized as follows. In the second section, the general proposed framework for forecasting the clearness index is presented; it is built by using different intervals for divide the observations taking into account the most signficant variable of the analyzed variables and by using different techniques for each interval. This section also briefly explains the statistical and machine-learning models that are integrated in the framework. In the third section, the data sets that have been used to fit the different used models are described. In this section, the different types of situation that the framework has to integrate are also analysed. In the fourth section, the results of forecasting the clearness index are presented; The results obtained when only one model is used vs the results obtained when different models are used for each set of observations are also analyzed in this section. Finally, the conclusions of the paper are summarized in the last section.

2 FRAMEWORK FOR THE SHORT-TERM FORECASTING OF THE CLEARNESS INDEX

We have developed a framework that allows the best model for short term forecasting of clearness index to be automatically selected, taking the actual value of this parameter into account.

Forecasting a time series is usually performed with a single model - statistical model or machine-learning model -, and using the recent past values of the series. However, this does not tally with the behavior of many types of phenomena: economic, climatic, etc. It would sometimes be very useful if the model used could be different for different times, taking into account that the influence of previous values can be different for different situations. Specifically, in the case of clearness index values, the type of relationship of the current value with the previous ones depends on what the current value is.

For example, if the last atmospheric clearness index values are very high (without clouds), there will be a high probability that the following value will also be high: this prediction can be only be performed using the most recent values of the series (usually one or two previous values are enough). Otherwise, if the previous values have varied greatly (patchy cloud), it would be necessary to use more values from the past (several hours and even previous days) to be able to predict the next value. Moreover, only some of these past values placed at any distance from the current value could be useful. It will then be necessary to determine what previous values have influence on the current value and to build a model that pick up these different relationships. As the previous values that influence a current value could be different, it would be very useful to be able to incorporate different models in the short term forecasting framework.

The first step to build the proposed framework has been data exploration in order to identify the most significant independent variable for forecasting (preliminary feature selection). We have analyzed the relationship of the current value with the previous ones and with other variables that affect this value such as the season of the year and the previous daily clearness index values. For this task, we propose the use of a multivariate regression. As a result of this first step, the most significant variable is selected. Several intervals are created on the significant variable in order to apply different models to each interval. In the second step,the model building and validation for the clearness index short term forecasting is addressed. We have analyzed what is the best short-term forecasting model for each set of observations from among the three different proposed models. These models are a multilayer neuronal network, a special type of probabilistic finite automata and a multivariate regression model. The use of each one of these model will depend on the results obtained when fitting these models for each interval. Once the framework is built by integrating all the fitted models, the short term forecasting is performed in an automatic way by using as input data the current value of the clearness index, the previous values of this parameter and the season of the year.

2.1 Statistical models

In the first step, we propose the use of a multivariate regression to determine what is the most significant variable for the clearness index series. The purpose of multivariate regression is to establish a quantitative relationship between a group of predictor variables (previous clearness index values and other variables that are the independent variables) and a dependent variable (the current value of the clearness index). The analyzed relationship is the following:

$$(Y_t) = f(Y_{t-1}, Y_{t-2}, ..., Y_{t-p}, S_1, S_2, S_3, S_4) \tag{2}$$

where $t-1, t-2, ...t-p$ are the previous hours from t and $\{S_i\}_{i=1}^{4}$ are the dummy variables to store the season of the year -a dummy variable is binary variable that takes the values 0 or 1 to indicate the absence or presence of some categorical effect; in this case, the dummy is used to indicate whether the observation happened or not in this season. Actually, in order to avoid

multicollinearity problems, only 3 dummy variables are included (dummy variables are used for spring, summer and autumn, no dummy variable is included for winter)

The results of this multiple regression analysis allow us to:

1. Understand which values in each situation have the greatest effects on the current value
2. Use the best model for forecasting clearness index values from the previous values that affect the current situation.

Several intervals are created on the significant variable in order to apply different models to each interval.

2.2 Machine Learning Models

We have implemented two machine learning models that are based on nonlinear time series modelling techniques:

1. Multilayer neural network
2. Probabilistic finite automata

Neural network models have been widely used in many practical problems. A description of these models can be found, for instance, in Hertz [1991], Hassoun [1995] and Anderson [1995]. We have used an artificial neural network with one hidden layer. Backpropagation learning algorith is implemented using the Laveberg-Marquad method. We have selected the $arctan$ function, as the transference function.

The probabilistic finite automata used is a special type of automata that allows some past values of the series to be learnt and others to be forgotten, as it is described in Mora-Lopez [2010]. It is based on the automata proposed in Ron [1996] but it allows information of different types to be incorporated, not only from the time series. This automata is capable of learning different independent variables for each interval of observations.

3 DATA SET

The data set that we have used are sequences of hourly global solar radiation from 10 Spanish locations. The total period of time that is used depends on the location; the total number of available monthly time sequence is 745. These data have been used to estimate the hourly clearness index by using the expression 1.

3.1 Data preparation

For the dependent variable hourly clearness index, k_h, d ($h = hour, d = day$), we have used the following independent variables:

1. The hourly clearness index values for the times: $h-1, h-2$ and $h-3$ the day d, that is $k_{h-1,d}, k_{h-2,d}, k_{h-3,d}$ that correspond to the three previous hours the same day.
2. The hourly clearness index values for the times: $d-1, d-2$ and $d-3$ the hour h, that is $k_{h,d-1}, k_{h,d-2}, k_{h,d-3}$ that correspond to the three previous days the same hour.

3. The season of the year for each observation. This information has been included with dummy variables, $S_1, S_2, S_3 and S_4$.

Only the values with all the previous variables defined (with values different from zero) are used in order to remove the hours that correspond to hours before sunrise and after sunset.

Moreover, the data have been filtered using the maximum observable value of hourly clearness. Each of the observations used for the experimentation has the following pattern:

$$(k_{h,d}, k_{h-1,d}, k_{h-2,d}, k_{h-3,d}, k_{h,d-1}, k_{h,d-2}, k_{h,d-3}, S_1, S_2, S_3) \tag{3}$$

where the first variable is the dependent variable.

The clearness index values are need to be discretized to build one of the machine learning models used, the probabilistic finite automata (PFA). We have used a static discrete conversion method. In general, the continuous values of a series Z_t are transformed into s discrete values through s intervals of same length. Specifically, the width w_X of a discretized interval is given by:

$$w_X = \frac{\max\{Z_t\} - \min\{Z_t\}}{s}, \tag{4}$$

where, hereafter, $\max$ and $\min$ are always considered for $t \in \{1, ..., n\}$. The discrete value z_i corresponding to a continuous value Z_i of the series is an integer from 1 to s which is given by:

$$z_i = discretize(Z_i) = \begin{cases} s & \text{if } Z_i = \max\{Z_t\} \\ [(Z_i - \min\{Z_t\})/w_X] + 1 & \text{otherwise} \end{cases}, \tag{5}$$

where $[A]$ means the integer part of A. After deciding upon s and finding w_Z, it is straightforward to transform the continuous values into discrete ones using this expression. We have used $s = 20$ (intervals), $max\{k_{h,d}\} = 1.0$, $min\{k_{h,d}\} = 0.0$ and the expression 5 to discretize the atmospheric clearness index values.

The observations in each interval are randomly divided into two set: the training set used for fitting the models and the test set using for validating the models. The first set includes the 90 per cent of the observations while the second one includes the 10 per cent.

4 RESULTS

In the first step, the exploratory analysis, the following multivariate regression was performed:

$$k_{h,d} = a_0 k_{h-1,d} + a_1 k_{h-2,d} + a_2 k_{h-3,d} + a_3 k_{h,d-1} + a_4 k_{h,d-2} + a_5 k_{h,d-3} + \sum_{j=6,i=1}^{j=8,i=3} a_j S_i \tag{6}$$

As expected, the most significant variable is $k_{h-1,d}$. Using the value of this variable, the training and set has been divided in the following 9 intervals using the following intervals (the last interval is greater as there are only a few values up to 0.85):

$$[0.0-0.1],]0.1-0.2],]0.2-0.3],]0.3-0.4],]0.4-0.5],]0.5-0.6],]0.6-0.7],]0.7-0.75],]0.75-1.0]$$

For the observations of each interval the three following models have been fitted: multilayer neural network, probabilistic finite automata and multivariate regression.

Table 1: Selected variables for each interval (*significance level=0.05*)

Interval for $k_{h-1,d}$	Selected var
0-0.1	$k_{h-1,d}$, $k_{h-3,d}$
0.1-0.2	$k_{h-1,d}$, $k_{h-3,d}$
0.2-0.3	$k_{h-1,d}$, $k_{h-3,d}$, $k_{h,d-3}$
0.3-0.4	$k_{h-1,d}$, $k_{h-3,d}$, $k_{h,d-1}$, $k_{h,d-3}$
0.4-0.5	$k_{h-1,d}$, $k_{h-3,d}$ $k_{h,d-1}$, $k_{h,d-3}$
0.5-0.6	$k_{h-1,d}$, $k_{h,d-1}$, $k_{h-3,d}$, $k_{h,d-3}$
0.6-0.7	$k_{h-1,d}$, $k_{h,d-1}$, $k_{h,d-2}$, $k_{h,d-3}$, S_3, $k_{h-3,d}$, S_1, S_2
0.7-0.75	$k_{h-1,d}$, $k_{h,d-1}$, $k_{h-2,d}$, $k_{h,d-2}$, S_3, S_2, $k_{h,d-3}$, S_1, $k_{h-3,d}$
0.75-1	$k_{h-2,d}$, $k_{h,d-1}$, $k_{h,d-3}$, $k_{h,d-2}$

An artificial neural network with one hidden layer using backpropagation learning algorithm, which perfors 10 epochs, is used. For training the neural network we have used all the independent variables, both the variables that represent the previous values in the series and the dummy variables for including the season of the year. The multivariate regression estimated for each observation set is the expression 6. Only the significant variables in each interval have been included to build the PFA. Table 1 reports the significant variables for each set.

Two measures of accuracy were estimated to assess the performance of each model:: the mean square error (MSE) and the relative error (RE). The expressions used are the following:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\frac{(k_t^*-k_t)^2}{k_t^2} \quad \text{RE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|(k_t^*-k_t)|}{k_t} \tag{7}$$

Where N are the number of observations and k_t^* is the estimated value of clearness index for the value k_t.

Figure 1 and Table 2 show the mean square error and relative error obtained for the clearness index short-term forecasting for each interval and each model for both the training set and the test set. As is shown, the model with the lowest MSE and RE for each interval is not always the same.

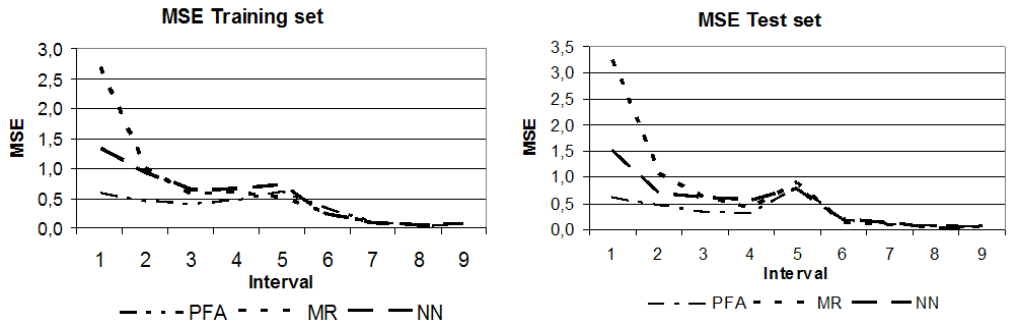

Figure 1: MSE for the the different models PFA: probabilistic finite automata, MR: multivariate regression, NN: neural network. Training and test sets.

The forecasting framework includes the model with the lowest RE (see Table 2 for each interval. As can be observed, the integration of different models for different intervals of observations allows better predictions than when the same model is used for all the data. Prediction can be improved by up to 20 % for some intervals.

Tables 2 and 3 set out the results obtained. As it can be observed, the integration of different models for different intervals of observations allows better predictions than when the same model is used for all the data. That is, the behaviour of the analyzed parameter (the clearness index) can be better modelled using different techniques depending on the current value; this means that the short term forecasting will be more accurate if it is possible to use the best model for each situation. In the training set, the average MSE decreases by more that 30 percent when using the best model when comparing with PFA, MR and NN models; in the test set, the MSE decreases between 30 and 50 per cent depending on the model. Improvement in prediction accuracy are also observed regarding the relative error when the results obtained for model integrated in the framework are compared with any other of three checked models.

Table 2: Relative error (eq.7) of the fitted models for the training and test sets. PFA: probabilistic finite automata, MR: multivariate regression, NN: neural network.

	Training set			Test set		
Interval	PFA	MR	NN	PFA	MR	NN
1	0,524	1,044	0,709	0,552	1,137	0,760
2	0,431	0,570	0,531	0,444	0,581	0,505
3	0,410	0,444	0,455	0,380	0,416	0,468
4	0,376	0,383	0,396	0,379	0,382	0,419
5	0,367	0,330	0,341	0,382	0,352	0,324
6	0,265	0,233	0,231	0,239	0,207	0,224
7	0,117	0,109	0,105	0,113	0,108	0,113
8	0,068	0,069	0,066	0,075	0,075	0,064
9	0,079	0,083	0,081	0,095	0,086	0,083
mean	0,208	0,224	0,213	0,206	0,223	0,215

Table 3: Estimated average MSE and RE for each model and for the model integrates in the framework (best model) for all observations.

	Training set		Test set	
Model	MSE	RE	MSE	RE
PFA	0,299	0,208	0,284	0,206
MR	0,310	0,224	0,419	0,223
NN	0,293	0,213	0,291	0,215
Best model (PFA,MR,NN)	0,203	0,197	0,199	0,193

5 CONCLUSIONS

We propose a framework for short-term forecasting of the clearness index based in the selection of the best model for each situation of this climatic parameter. The built framework integrates both statistical and machine learning models. A multivariate regression was used to identify the different relationships observed in the clearness index series. This regression allows us to select the most significant variable in the series and to divide the observations in several intervals. For each of these intervals, we checked three different models for short-term forecasting: a multivariate regression, a special type of probabilistic finite automata and an artificial neural network.

In the developed framework, the short-term forecasting best model for each interval is integrated. Once the framework is built, it is possible to predict the next value of clearness index only using the current and previous values of this parameter and the season of the year. The integration of different models for different situations has to allow an improvement greater that 30 per cent in short-term forecasting with respect to when a single model is used.

ACKNOWLEDGMENTS

This work has been partially supported by the projects TIN2008-06582-C03-03, P07-RNM-02504 and ENE07-67248 of the Spanish Ministry of Science and Innovation (MICINN).

REFERENCES

Aguiar, R., C.-P. M.-C. J. Simple procedure for generating sequences of daily radiation values using a library of markov transition matrices. *Solar Energy*, 40(3):269–279, 1988.

Aguiar, R. and M. Collares-Pereira. A.g: A time dependent autoregressive gaussian model for generating synthetic hourly radiation. *Solar Energy*, 49(3):167–174, 1992.

Al-Alawi, S.M., A.-W.-S. B. C. Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone. *Environmental Modelling & Software*, 23(4):396–403, 2008.

Anderson, J. A. *An Introduction to Neural Networks*. The MIT Press, USA, 1995.

Bartoli, B., C. B.-C.-V. F. M.-S.-C. Autocorrelation of daily global solar radiation. *Il nuovo cimento*, 40:113–122, 1983.

Box, G. and G. Jenkins. *Time Series Analysis forecasting and control*. Prentice Hall, USA, 1976.

Brinkworth, B. Autocorrelation and stochastic modelling of insolation sequences. *Solar Energy*, 19:343–347, 1977.

Cao, J. and X. Lin. Study of hourly and daily solar irradiation forecast using diagonal recurrent wavelet neural networks. *Energy Conversion and Management*, 49(6):1396–1406, 2008.

Elminir, H.K., A. Y.-Y.-F. Prediction of hourly and daily diffuse fraction using neural network, as compared to linear regression models. *Energy*, 32(8):1513–1523, 2007.

Ghanbarzadeh, A., N. A.-A.-E. e. a. Solar radiation forecasting based on meteorological data using artificial neural networks. In *7TH IEEE International Conference on Industrial Informatics, VOLS 1 AND 2*, pages 227–231, 2009.

Graham, V. A., H.-K. G. T. U.-T. E. A time series model for kt with application to global synthetic weather generation. *Solar Energy*, 40:83–92, 1988.

Hassoun, M. H. *Fundamentals of Artificial Neural Networks*. The MIT Press, USA, 1995.

Hertz, J., K. A.-P. R. G. *Introduction to The Theory of Neural Computation*. Addison-Wesley Publishing Company, USA, 1991.

Ingsrisawang, L., I. S.-S. S. A. P.-K. W. Machine learning techniques for short-term rain forecasting system in the northeastern part of thailand. In *Proceedings of World Academy of Science, Engineering and Technology. V. 31, July 2008*, pages 248–253, 2008.

Kilsby, C.G., J. P.-B. A. F. A.-F. H. H. C. J. P. S. A. W. R. A daily weather generator for use in climate change studies. *Environmental Modelling & Software*, 22 (12):1705–1719, 2007.

Mora-Lopez, L., M.-J. P. M. S.-d.-C. M. An intelligent memory model for short-term prediction: an application to global solar radiation data. In *Twenty Third International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems*, pages 1–10, 2010.

Mora-Lopez, L., S.-d.-C. M. Multiplicative arma models to generate hourly series of global irradiation. *Solar Energy*, 63:283–291, 1998.

Mubiru, J., B. E. Estimation of monthly average daily global solar irradiation using artificial neural networks. *Solar Energy*, 82(2):181–187, 2008.

Ron, D., S. Y. T.-N. The power of amnesia: Learning probabilistic automata with variable memory length. *Machine Learning*, 25:117–149, 1996.