Modeling time series of climatic parameters with probabilistic finite automata

L. Mora-López a,*, J. Mora b, R. Morales-Bueno a, M. Sidrach-de-Cardona c

a Dpto. Lenguajes y C. Computación, E.T.S.I. Informática, Universidad de Málaga, Campus Teatinos, 29071 Malaga, Spain
b Dpto. Fundamentos del Análisis Económico, Universidad de Alicante, Apdo. Correos 99, E-03080 Alicante, Spain
c Dpto. Física Aplicada II, E.T.S.I. Informática, Universidad de Málaga, Campus Teatinos, 29071 Malaga, Spain

Received 1 November 2002; received in revised form 22 October 2003; accepted 20 April 2004

Abstract

A model to characterize and predict continuous time series from machine-learning techniques is proposed. This model includes the following three steps: dynamic discretization of continuous values, construction of probabilistic finite automata and prediction of new series with randomness. The first problem in most models from machine learning is that they are developed for discrete values; however, most phenomena in nature are continuous. To convert these continuous values into discrete values a dynamic discretization method has been used. With the obtained discrete series, we have built probabilistic finite automata which include all the representative information which the series contain. The learning algorithm to build these automata is polynomial in the sample size. An algorithm to predict new series has been proposed. This algorithm incorporates the randomness in nature. After finishing the three steps of the model, the similarity between the predicted series and the real ones has been checked. For this, a new adaptable test based on the classical Kolmogorov–Smirnov two-sample test has been done. The cumulative distribution function of observed and generated series has been compared using the concept of indistinguishable values. Finally, the proposed model has been applied in several practical cases of time series of climatic parameters.

© 2004 Elsevier Ltd. All rights reserved.

Keywords: Machine learning; Modeling climatic data; Time series

Software availability

Name: SIMULTS
Developer: L. Mora-López
Contact address: llanos@lcc.uma.es, Tel.: +34-95-213-2802; fax: +34-95-213-1397
Year first available: 2002
Hardware required: Pentium II (minimum)
Software required: None
Program language: C
Program size: 314 K
Cost: Free

1. Introduction

The fundamental idea in this paper is the use of probabilistic finite automata (PFA) as a means of representing the relationships observed in stationary time series. PFA are mathematical models used in the machine-learning field. Traditionally, the analysis of stationary time series has been carried out using stochastic process theory. One of the most detailed analyses of statistical methods for time series research was done by Box and Jenkins (1976). The goal of these methods is to find models which are able to reproduce the statistical and sequential characteristics of the series. Usually, the approach is as follows (see Box and Jenkins, 1976): first, the recorded series are statistically analyzed in order to select the best model for the series. Then the parameters of the model are estimated. After this, a new series of values can be generated using the estimated model. For example, for solar radiation series
This approach has been followed in Brinkworth (1977), Bendt et al. (1981), Aguiar et al. (1988), Aguiar and Collares-Pereira (1992), and Mora-López and Sidrach-de-Cardona (1997). Many of these models can only generate new sequences of values which present normal probability distribution functions. However, the original series do not have this type of probability distribution. Other types of time series which have been analyzed using some class of stochastic models are shown in McMillan et al. (2000) and Anh et al. (1997).

On the other hand, the analysis of time series and stochastic process has also been analyzed from machine-learning techniques. When a time series presents a probabilistic behavior, some machine-learning models could be very useful to study it. In these series the recorded variables are insufficient to exactly determine the future values, due to the random nature of these variables. The systems in which these models can be used must have the following properties:

To present probabilistic behaviour or uncertainty.
This uncertainty can be due to several factors. For example, for the prediction of climatic variables the number of parameters which affect them is very high. Although there is uncertainty in these systems, there is always some structure within this uncertainty.

For example, the machine-learning models based on probabilistic finite automata have been used to model several types of natural sequences. Examples of such applications are: universal data compression (Rissanen, 1983), analysis of biological sequences, for DNA and proteins (Krog et al., 1993), analysis of natural language, for handwriting and speech (Nadas, 1984; Rabiner, 1994; Ron et al., 1998), etc. Different classes of automata have been developed. For instance, acyclic probabilistic finite automata have been used for modeling distributions on short sequences (Ron et al., 1998); probabilistic suffix automata, based on variable order Markov models, have been used to construct a model of the English language (Ron et al., 1994). All these automata allow us to take into account the temporal relationships in a series. In a different way, other approaches from machine learning have been used to model climatic parameters. For instance, neural networks have been used by Mohandes et al. (1998) and Kemmoku et al. (1999) on the characterization of one climatic parameter (daily global solar radiation). The main problem of this approach is that the obtained models are ‘black boxes’, and no significant information can be obtained from them.

Other works arise from the ideas developed by Dagum, based on belief network models: it is proposed the use of dynamic network models, which are a compromise between belief network models and classical models of time series. They are based on the integration of fundamental methods of Bayesian analysis of time series. This paper describes how to use certain models from the machine-learning field in the analysis and prediction of climatic parameters. The model we propose is based on the Probabilistic Finite Automata (PFA) theory. Our goal is to use PFA to represent all the relationships observed in stationary climatic time series and to use these PFA to predict new values of the series. The use of this model allows us to represent and generate time series with non-normal probability distribution functions and to obtain information about the nature of the analyzed series. Moreover, an adaptable test based on the classic Kolmogorov–Smirnov two-sample test has been used to check the proposed model. Finally, preliminary results of the model obtained for climatic parameters are shown.

2. Probabilistic finite automata

We propose using a mathematical model called probabilistic finite automata (PFA). We propose the use of this mathematical model to represent a stationary univariate time series. Formally, a PFA is a 5-tuple \((\Sigma, Q, \tau, \gamma, q_0)\) where (see for instance, Ron et al. (1998) or Mora-López and Sidrach-de-Cardona (2003)):

- \(\Sigma\) is a finite alphabet; that is, a set of discrete symbols corresponding to the different continuous values of the analyzed parameter. The different symbols of \(\Sigma\) will be represented by \(x_i\). For a series, the values observed can be \(x_1, x_2, \ldots, x_3\). To represent the different observable series for a period \(t_1\) to \(t_m\) we will use the symbols \(y_{i_1}, y_{i_2}, \ldots, y_{i_m}\). So, in the series \(x_1, x_2, \ldots, x_3\), the symbol \(y_1\) corresponds to the value \(x_5\), the symbol \(y_2\) to \(x_6\) and so on.
- \(Q\) is a finite collection of states. Each state corresponds to a subsequence of the discretized time series. The maximum size of a state—number of symbols—is bounded by a value \(N\) fixed in advance. This value is related to the number of previous values which will be considered to determine the next value in the series and depends on ‘memory’ of the series. The ‘memory’ of one series could be estimated both empirically and using information about the parameter which is analyzed. In the first case, an iterative process could be used: a small value is selected at the beginning; this value is increased in each iteration, and the results obtained with the previous value of memory and the actual value are compared; the process continues until the models do not improve when the memory increases. In the second case, the previous information about the analyzed parameter could be used to initialize the value of the memory in the iterations.

- \(\tau: Q \times \Sigma \rightarrow Q\) is the transition function
- \(\gamma: Q \times \Sigma \rightarrow [0, 1]\) is the next symbol probability function
- \(q_0\in Q\) is the initial state

The function \(\gamma\) satisfies the following requirement: For every \(q\in Q\) and for every \(x_i\in \Sigma\), \(\sum_{i=1}^{N} \Sigma \gamma(q, x_i) = 1\). Moreover, the following conditions are required:

\[
\tau: Q \times \Sigma \rightarrow Q
\]

\[
\gamma: Q \times \Sigma \rightarrow [0, 1]
\]

\[
q_0\in Q
\]
The transition function \( \tau \) can be undefined only on states \( q \in Q \) and symbols \( x \in \Sigma \), for which \( \gamma(q,x) = 0 \).

The function \( \tau \) can be extended to be defined on \( Q \times \Sigma^* \) in the following recursive manner:

\[
\tau(q,y_1,y_2,\ldots,y_t) = \tau(\tau(q,y_1,y_2,\ldots,y_{t-1}),y_t)
\]

where \( y_i \in \Sigma \).

Graphically, each state is represented by a node and the edges going out of each state are labeled by symbols drawn from the alphabet. Moreover, each state has an associated probability vector which is composed of the probability of the next symbol for each of the symbols of the alphabet.

For instance, in Fig. 1 a simple PFA is shown. In this PFA, the alphabet, \( \Sigma \), is composed of the symbols 0 and 1. The states of the system, \( Q \), are described in each node of the automata: initial (i), 0, 1, 00, 01, 10, and 11. For instance, the state labeled 01 corresponds to the following sequence of values in the series: 1 as the last value and 0 as the previous. The associated vectors at each state (node) are the probabilities which each symbol of the alphabet has to appear in the next moment, after the sequence of symbol that label the node has appeared. For instance, the node labeled with 10, has the associated vector \((0.25,0.75)\); this means that if the current state is 10, then the next symbol can be 0, with a probability of 0.25 and 1 with a probability of 0.75. The continuous and discontinuous arrows represent the transition function between states (discontinuous for 0, and continuous for 1). For instance, if the current state is 10, and the next symbol is 0, then the following state will be labeled with 00; but if the next symbol is 1, then the following state will be labeled with 01.

In the PFA shown in Fig. 1, the states 01 and 11 have the same probability vector as state 1. That is, when the symbol 1 appears, it is not necessary to know the preceding value to determine the probabilities of the next symbol, since in both cases (0 or 1), the probabilities vector of the next symbol is \((0.5,0.5)\). Therefore, the PFA of Fig. 1 can be converted into the PFA shown in Fig. 2.

This class of PFA is used to represent variable order Markov models. These simplified automata are the automata proposed in this paper. They capture the same information with fewer states than the original automata. Moreover, they allow us to take into account, for each state, a different number of previous values in the series.

Let us define some concepts that we will use to build the PFA for climatic data series. Let \( \Sigma = \{x_1,x_2,\ldots,x_n\} \) be the set of discrete values of the analyzed variable and \( \Sigma^* \) denote the set of all possible sequences which can be obtained with these values. For any integer \( N \), \( \Sigma^N \) denotes the set of all possible sequences of length \( N \), and \( \Sigma \leq N \) is the set of all possible sequences with length less than or equal to \( N \). For any subsequence, \( Y \), represented by \( y_1\ldots y_m \), where \( y_i \in \Sigma \), the following notations will be used:

- The longest final subsequence of \( Y \), different from \( Y \), will be \( \text{final}(Y) = y_2\ldots y_m \)
- The set of all final subsequences of \( Y \) will be \( \text{last}(Y) = \{y_i\ldots y_m|1 \leq i \leq m\} \)

In the next section we explain how to build a PFA for a time series.

3. Building probabilistic finite automata

3.1. Algorithm to build probabilistic finite automata

The following algorithm is used to build the PFA:

- **Step 1.** Compute the series of discrete values.
- **Step 2.** Initialize the PFA with a node, with label null sequence.
- **Step 3.** The set \( PSS — \text{Possible Subsequence Set} \) — is initialized with all sequences of order 1. Each element in this set corresponds to a sequence of discrete values. Take \( o = 1 \) as the initial value of the order, i.e. size of subsequences to consider.
- **Step 4.** If there are elements of order \( o \) in \( PSS \), pick any of these elements, \( Y \). Using all discrete sequences in the series, compute the frequency of \( Y \). If 4a and 4b are true, then go to step 5, else go to step 6.
- **Step 4a.** The frequency of this sequence is greater than the threshold frequency. The value of this frequency depends on the size of the sample
used to build the PFA. The more available data there are, the greater this threshold frequency can be. This threshold value is used to ensure that the model is not overfitted with the sample.

Step 4b. For some \( x_p \in \Sigma \), the probability of occurrence of the subsequence \( Yx_p \) is not equal to the probability of the subsequence \( \text{final}(Y)x_p,s \), that is: \( P(x_p|Y) \neq P(x_p|\text{final}(Y)) \) (not equal: when the ratio between the probabilities is significantly greater than one).

Step 5. Do
Step 5a. Add to the PFA a node, labeled with \( Y \), and compute its corresponding probabilities vector.
Step 5b. For each amplified sequence, \( Yx_p \): if the probability of this augmented sequence is greater than the threshold probability, then include it in \( \text{PSS} \).
Step 6. Remove the analyzed subsequence, \( Y \), from \( \text{PSS} \).
Step 7. If there are no elements of order \( o \) in \( \text{PSS} \), add 1 to the value of \( o \). If \( o \leq N \) and there are elements of length \( o \) in \( \text{PSS} \), then go to step 4, else Stop.

3.2. Predicting new values

A PFA can be used as a mechanism for generating finite sequences of values in the following manner. Start from an initial value selected from the alphabet, called the initial state. If \( q_i \) is the current state, labeled by the sequence \( Y = y_1...y_t \), then the next symbol is chosen (probabilistically) according to \( \gamma(q_i) \). If \( x \in \Sigma \) is the chosen symbol, then the next state, \( q_{i+1} \), is \( \tau(q_i,x) \). The label of this new state, \( Y' \), will be the longest final subsequence of \( Yx \) in the PFA, that is:

\[
Y' = \text{Max}\{\text{last}(Yx)\} \epsilon \text{PFA}
\]

The process continues until the length of the required sequence is reached.

Moreover, if \( P'(Y) \) denotes the probability that a PFA generates a sequence \( Y = y_1...y_{t-1}y_t \), then:

\[
P'(Y) = \prod_{i=0}^{t-1} \gamma(q_i,y_{i+1})
\]

This definition implies that \( P'(Y) \) is in fact a probability distribution over the symbols of sequence, i.e.:

\[
\sum_{Y \in \Sigma} P'(Y) = 1
\]

4. How the model can be validated

For a recorded time series, the following steps must be followed to use the proposed model. First, if the time series has continuous values, then these values must be discretized. After this, the PFA is built using the discrete series. With the PFA and the generation method described above, new values for the time series can be generated.

In order to compare the simulated series to the real ones, several statistical tests can be used. The hypothesis that both series have the same mean and variance has been checked. The frequency histograms of the recorded and simulated series are also analyzed.

To make this comparison, we propose the use of an adaptable goodness-of-fit test, which is based on the two-sample Kolmogorov−Smirnov test, described in Rohatgi (1976); we have used this test—instead of an ANOVA test for the mean or Levene’s test for the variance—because it captures both the differences in mean and variance and the differences in any characteristic of the probability distribution function. The objective of this adaptable test is to determine if two distribution functions \( F_Y(.) \) and \( F_{Y'}(.) \) are the same, except for possible changes in location and scale. Specifically, we have checked the null hypothesis that there exist two unknown values \( \mu \) and \( \sigma \) such that \( Z_t \) and \( \mu + \sigma Y_t \) have the same distribution.

Using distribution functions it is possible to express our null hypotheses as follows:

\[
H_0 : \exists \mu \epsilon \mathbb{R}, \sigma \in (0, + \infty) / \forall \mu \epsilon \mathbb{R}
\]

\[
F_X(u) = F_Y\left(\frac{u - \mu}{\sigma}\right)
\]

Replacing unknown parameters \( \mu \) and \( \sigma \) by estimates introduces additional random terms in the statistic and traditional critical values cannot be used. Therefore, to obtain the critical values that must be used in the test, we propose using a bootstrap procedure.

5. Practical cases: using probabilistic finite automata for climatic data

The probabilistic finite automata presented in the previous sections have been used to characterize and predict two climatic variables: the hourly global radiation received on a surface on the ground and the hourly dry bulb temperature. For these variables, time series are recorded by meteorological stations at regular time intervals. In order to evaluate the effect of using information about the series, such as trend, probability distribution of the parameter, and so on, for the hourly global radiation series we have used some previous information about the series, but for the hourly dry bulb temperature series we do not have used any information about the nature of this parameter. In Sections 5.3 and 5.4 we explain what information has been used.
As it has been mentioned above, we need stationary time series. For each series we have used a different method to obtain a stationary series, as is explained in the corresponding sections (Section 5.3 for hourly global radiation and Section 5.4 for dry bulb temperature).

The following question—which we have solved—is the discretization of these series. The recorded values are continuous whereas the proposed mathematical model uses discrete values. The discretization method used is explained later. The PFA have been built using the discrete series obtained and new values of the series generated. Finally, we have checked these values using several tests.

5.1. Data set

The data of the hourly exposure series of global radiation, \( \{G_h(t)\} \), which are used in this work were recorded over several years in nine Spanish meteorological stations. The used stations (latitude) are: Málaga (36.66N), Sevilla (37.42N), Murcia (38.00N), Badajoz (38.89N), Mallorca (39.33N), Castellón (39.95N), Madrid (40.45N), Tortosa (40.81N) and Oviedo (43.35N).

The pertinent latitudes range from 29.32N to 39.18N, and are used are included in the Solar and Meteorological Surface Observation Network from the U.S. Department of Energy of the National Renewable Energy Laboratory. The used stations (latitude) are: San Antonio (29.32N), Port Arthur (29.57N), Lake Charles (30.7N), Baton Rouge (30.32N), San Angelo (31.22N), Jackson (32.19N), Little Rock (34.44N), Amarillo (35.14N), Springfield (37.14N), Evansville (38.2N), Columbia (38.49N) and Kansas City (39.18N).

The pertinent latitudes range from 29.32N to 39.18N, and the years range from 1961 to 1990. In total, 4294 months were accounted for (only months without missing data are used).

The data of dry bulb temperature which have been used are included in the Solar and Meteorological Surface Observation Network from the U.S. Department of Energy of the National Renewable Energy Laboratory. The used stations (latitude) are: San Antonio (29.32N), Port Arthur (29.57N), Lake Charles (30.7N), Baton Rouge (30.32N), San Angelo (31.22N), Jackson (32.19N), Little Rock (34.44N), Amarillo (35.14N), Springfield (37.14N), Evansville (38.2N), Columbia (38.49N) and Kansas City (39.18N).

The pertinent latitudes range from 29.32N to 39.18N, and the years range from 1961 to 1990. In total, 4294 months were accounted for (only months without missing data are used).

5.2. Discretization of the series

The goal is to use an effective and efficient method to transform continuous values into discrete ones using the overall information included in the series and, when possible, feedback with the learning system. To do this, the discrete value which corresponds to a continuous value has been calculated using qualitative reasoning, taking into account the evolution of the series. Qualitative reasoning models have been used in different areas in order to obtain a representation of the domain based on properties (qualities) of the systems; see, for instance, Forbus (1984), Kleer and Brown (1984) and Kuipers (1984).

We have used a qualitative dynamic discrete conversion method, described in Mora-López et al. (2000). It is dynamic because the discrete value associated to a particular continuous value can change over time: that is, the same continuous value can be discretized into different values, depending on the previous values observed in the series. It is qualitative because only those changes which are qualitatively significant appear in the discretized series. The number of states of the built PFA is sensitive to the discretization method used, as Mora-López et al. (2000) explain.

5.3. PFA for global hourly solar radiation series

The series of hourly global radiation have daily and monthly trends due to the position of the sun. Using this information about these series, we have used the series of the hourly clearness index calculated from the hourly global radiation. These estimated series are stationary. The hourly clearness index is defined as:

\[
K_h = \frac{G_h}{G_{h,0}}
\]

where \( G_h \) is the hourly global radiation and \( G_{h,0} \) is the extraterrestrial hourly global radiation.

First, it is necessary to determine the size of the alphabet which will be used to build the PFA. We have analyzed the monthly probability distribution functions, pdf, of the clearness index in order to determine how many intervals—that is, how many different symbols—will be used to classify the continuous values of this parameter—which ranges between 0 and 1. Taking into account the total number of months which are available to build the PFA and the estimated pdf, eight different intervals have been used. The alphabet proposed for the PFA is:

\[
\Sigma = \{0, 1, \ldots, 7\}
\]

The relationship between the values of the clearness index and the symbols of the alphabet has been determined by using the probability distribution function of the series. For the first symbol of a series, the discrete value of the series will be calculated using:

\[
Y_h = \begin{cases} 
0 & 0 \leq K_h < 0.35 \\
\left\lfloor \frac{K_h - 0.35}{0.05} \right\rfloor + 1 & 0.35 \leq K_h < 0.65 \\
7 & K_h \geq 0.65
\end{cases}
\]

where \( \lfloor A \rfloor \) means the integer value of \( A \). For the following values of the continuous series the discretized values will be obtained using the algorithm described in Mora-López et al. (2000) and the aforementioned expression.

Using these expressions and the hourly clearness index series, the discrete series \( \{Y_h\} \) are obtained. From all possible subsequences observed in the series, only those
with a sufficient probability will be used to build the PFA. This threshold of probability must be defined when the PFA is built—as we have already explained in Section 3.1.

Finally, the monthly series of the hourly clearness index have been grouped using the monthly mean value of the hourly clearness index. The ranges for each group are the same as those defined for the discretization of this parameter. For every interval, one PFA has been built.

5.4. PFA for dry bulb temperature series

We have transformed the original dry bulb temperature series into stationary series by using the hourly mean and variance of the monthly series; thus, no additional information about the parameter has been used. For the data of each month $m$, the modified series has been obtained using the expression:

$$ D_{h,m} = \frac{\sum_{d=1}^{nyears} \sum_{m=1}^{\text{month}_m} D_{h,d,m,y}}{\text{ndays}_m * \text{month}_m} $$

where $D_{h,d,m,y}$ is the original values of the time series for hour $h$, day $d$, month $m$ and year $y$, $D_{h,m}$ is the mean of the observed values for hour $h$ and month $m$, and $\text{var}_{h,m}$ is the variance. These two values are calculated using all the data available for hour $h$ and month $m$ across the years. The values of $h$ range between 1 and 24, and the values of $m$ range between 1 to 12. The expression which has been used is:

$$ X_{h,d,m} = \frac{D_{h,d,m,y} - D_{h,m}}{\text{var}_{h,m}} $$

Finally, in order to obtain series which range between $-1$ and 1, the values of the series have been divided by the value $|X_{h,d,m,y}|$ maximum, where $|X|$ represents absolute value.

The alphabet of the PFA is:

$$ \Sigma = \{0,1,\ldots,9\} $$

The relationship between the values of the normalized series and the symbols of the alphabet is the following. For the first symbol of a series, the discrete value of the series will be calculated using the expression:

$$ Y_h = 5 + |X_h * 5| $$

where $|A|$ means the integer value of $A$.

For the following values of the continuous series the discrete values will be obtained using the algorithm described in Mora-López et al. (2000) and the aforementioned expression.

Using these expressions and the dry bulb normalized series, the discrete series $\{Y_h\}$ are obtained. From all possible subsequences observed in the series, only those with a sufficient probability will be used to build the PFA. This threshold of probability must be defined when the PFA is built. As no additional information about the original series has been used for these series, only one PFA has been built using all the series.

5.5. Predicting new values

To generate new series we need an initial state. The initial state is the discrete value corresponding to the mean value of each time series—clearness index and normalized series of dry bulb temperature. Let $q_i$ be the current state. The next symbol, $y_{t+1}$, is generated as follows: first, a random number $r \in [0,1]$ is generated. Then, we choose the only component of probabilities vector—for the current state, $q_i$—which satisfies:

$$ y_{t+1} = x_j \left\{ \sum_{i=1}^{j} \gamma(q_i, x_i) \right\} r \quad \text{AND} \quad \sum_{i=1}^{j-1} \gamma(q_i, x_i) < r $$

5.6. Results

To select the best values of the parameters to build the PFA, we have checked the results obtained with different values of these parameters. For both climatic parameters, the values we have used are:

- Order of the PFA: from 2 to 14.
- Threshold—minimum number of appearances of a sequence—from 1 to 5 (these values have been selected taking into account the amount of samples used to build the PFA).

For the clearness index, for most of the intervals, if the order used for the PFA is 2, the results are similar to those when the order is 4; however, for intervals 5 and 6, using order 4, the PFA captures the relationship observed in the series better than using order 2. Thus, the selected order (maximum) for the PFA is 4. The selected minimum number of appearances—required to use a sequence to build a PFA—is 2.

### Table 1

<table>
<thead>
<tr>
<th>Interval</th>
<th>Months</th>
<th>Similar</th>
<th>Perc</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0.0, 0.35)</td>
<td>17</td>
<td>17</td>
<td>100</td>
</tr>
<tr>
<td>[0.35, 0.4)</td>
<td>55</td>
<td>54</td>
<td>98.2</td>
</tr>
<tr>
<td>[0.4, 0.45)</td>
<td>79</td>
<td>78</td>
<td>98.7</td>
</tr>
<tr>
<td>[0.45, 0.5)</td>
<td>107</td>
<td>106</td>
<td>99.1</td>
</tr>
<tr>
<td>[0.5, 0.55)</td>
<td>137</td>
<td>136</td>
<td>99.3</td>
</tr>
<tr>
<td>[0.55, 0.6)</td>
<td>198</td>
<td>192</td>
<td>97.0</td>
</tr>
<tr>
<td>[0.6, 0.65)</td>
<td>120</td>
<td>116</td>
<td>96.7</td>
</tr>
<tr>
<td>[0.65, 1.0)</td>
<td>32</td>
<td>30</td>
<td>93.8</td>
</tr>
</tbody>
</table>

The third column shows the number of months predicted as being similar to the real ones. The fourth column shows the percentage.
With the built PFA, new sequences of the hourly clearness index have been generated. The original and generated series have been compared using the statistical test described above.

The results obtained for each interval of the clearness index are shown in Table 1.

In Fig. 3, the cumulative probability distribution function of both series of clearness index—recorded and simulated—are shown (data from Madrid).

Moreover, we have calculated the hourly series of global irradiation from the hourly clearness index series. Using statistical tests, the null hypothesis that both series have the same mean and variance is not rejected (significance level \( \alpha = 0.05 \)). For instance, in Figs. 4 and 5, the recorded and simulated series of one Spanish location are shown (Malaga, January 1977).

The frequency histograms of the recorded and simulated series of hourly global radiation have been also analyzed. The frequency histograms have been obtained for each month of the year, using all the recorded and simulated series for that month over every year. The null hypothesis that the underlying model for both series is the same has never been rejected (significance level \( \alpha = 0.05 \)).

The results obtained for normalized dry bulb temperature series are slightly different to the ones obtained for the clearness index series. First, as we do not have used any additional information about the series, we have analyzed the dependence between the accuracy of the simulation and the order used for the PFA. In Table 2 is shown the number of months (in percentage) in which the null hypothesis that both series have the same mean, variance and probability distribution function is not rejected (significance level \( \alpha = 0.05 \)), as a function of the used order. As can be seen, the accuracy of the model depends more strongly on the order of the PFA than in the case of the clearness index series. A possible explanation for this is that the ‘memory’—order—of the PFA captures the information in the series that we have not used. However, the results do not reach the degree of accuracy achieved with the clearness index series, for which the model has been constructed using additional information about the parameter. In any case, the results obtained for the dry bulb temperature series prove that the PFA models can be used to represent and generate this time series. Although no information about the series has been used to estimate the model, when the order is 10 more than 90% of the simulated series are statistically similar to the real ones according to the results of the Kolmogorov–Smirnov test (significance level \( \alpha = 0.05 \)).

6. Conclusions

In this paper, a new model to predict climatic parameters is proposed. This model is based on the use of probabilistic finite automata and has been

<table>
<thead>
<tr>
<th>Order</th>
<th>% Similar months</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>79.5</td>
</tr>
<tr>
<td>4</td>
<td>80.5</td>
</tr>
<tr>
<td>5</td>
<td>82.3</td>
</tr>
<tr>
<td>6</td>
<td>84.7</td>
</tr>
<tr>
<td>7</td>
<td>85.2</td>
</tr>
<tr>
<td>8</td>
<td>86.9</td>
</tr>
<tr>
<td>9</td>
<td>88.5</td>
</tr>
<tr>
<td>10</td>
<td>90.2</td>
</tr>
</tbody>
</table>
developed within the machine-learning field. We have verified that this model allows us to keep all the relevant information contained in the univariate time series in an easy way. Moreover, with this mathematical model, the different relationships observed between different subsequences can be left out; each subsequence only uses the memory length that it requires.

Using this model, two series of climatic parameter - global hourly radiation data and dry bulb temperature - have been analyzed. For the first one, the model has been built using additional information about the analyzed parameter, such as the definition of intervals to discretize the series, the use of different models for different types of month or the use of other variables - solar extraterrestrial radiation - to get stationary series. For the second one, only information contained in the series has been used: the mean and the variance. To estimate the model, first a dynamic discretization method was used. Then, a set of PFA was built: for clearness index, one PFA for each interval of the analyzed parameter; for dry bulb temperature, only one PFA. Using these PFA, a method to predict new values of the parameter has been proposed.

Finally, the model has been checked using several tests. To check the cumulative probability distribution functions, an adaptable test, based on the two-sample Kolmogorov–Smirnov test, was used.

The obtained results prove that Probabilistic Finite Automata can be used to model climatic parameters and to predict new values in a suitable way. Moreover, if additional information about the time series is used, the results can improve significantly.

Acknowledgements

This work has been partially supported by MOISES Project, ref. TIC2002-04019-C03-02 and DGI project, ref. BEC-2002-03097, of the MCYT, Spain.

References


