The Box–Jenkins analysis and neural networks: prediction and time series modelling

Sana BuHamra a, Nejib Smaoui b,*, Mahmoud Gabr a

a Department of Statistics and Operations Research, Kuwait University, P.O. Box 5969, Safat 13060, Kuwait
b Department of Mathematics and Computer Science, Kuwait University, P.O. Box 5969, Safat 13060, Kuwait

Received 2 August 2001; received in revised form 10 March 2003; accepted 28 March 2003

Abstract

Two approaches, namely the Box–Jenkins (BJ) approach and the artificial neural networks (ANN) approach were combined to model time series data of water consumption in Kuwait. The BJ approach was used to predict unrecorded water consumption data from May 1990 to December 1991 due to the Iraqi invasion of Kuwait in August 1990. A supervised feedforward back-propagation neural network was then designed, trained and tested to model and predict water consumption from January 1980 to December 1999. It is interesting to note that the lagged or delayed variables obtained from the BJ approach and used in neural networks provide a better ANN model than the one obtained either blindly in blackbox mode as has been suggested or from traditional known methods.

© 2003 Elsevier Inc. All rights reserved.

Keywords: Box–Jenkins analysis; Autoregressive integrated moving average model; Artificial neural networks; Time series modelling

1. Introduction

Artificial neural networks (ANN) have been widely used to model time series in various fields of applications including dynamical systems [1–4], nonlinear signal processing [5], pattern recognition, identification and classification [6], speech [7,8], vision and control systems [9], and packet traffic [10]. The complexity observed and encountered in time series suggests the use of neural networks which have been proven to be capable of modelling complex nonlinear relationship
without a priori assumption of the nature of the relationship. Chaotic time series, for example, were modelled and predicted by controlling the nonlinearity in neural networks with prediction capability far exceeding conventional methods [2,5]. In general, prediction of the future state of a noncovariate time series is made by knowing the present measurements and possibly some recent history. Packard et al. [11] demonstrated that an attractor (i.e., a periodic trajectory, strange or chaotic attractors, fixed points, etc. . . ) may be reconstructed from a time series if a correct number of time delayed samples of the time series is used. The number of delayed samples was suggested by Takens [12] who proposed lower and upper bounds of the dimension of the sample space or embedding dimension $d_E$ (i.e., $d_F \leq d_E \leq 2d_F + 1$, with $d_F$ being the attractor fractal dimension). Following Packard and Takens ideas, Smaoui [2] modelled Henon attractors using neural networks where the fractal dimension $d_F$ was used as an indicator for the number of delayed variables needed in the input layer.

In this work, the total monthly water consumption in Kuwait from January 1980 to April 1990 and from January 1992 to December 1999 presented as time series are modelled (see Fig. 1). The total monthly water consumption presented in Fig. 1 in millions of gallons were obtained from the ministry of planning. Eventhough neither of the two time series presented is chaotic, predicting the future Kuwait water consumption remains a difficult task. Because of the Iraqi invasion of Kuwait, the water consumption data from May 1990 to December 1991 was not recorded. Therefore, if one needs to model the time series from 1980 to 1999, then the unrecorded data should first be predicted. The Box–Jenkins (BJ) approach is used with the task of predicting the missing data. Information regarding the appropriate number of delayed variables obtained from BJ analysis is then used in ANN.

The paper is organized as follows: In Section 2, the BJ approach on the water consumption data is presented. Using this approach, prediction of unrecorded data due to the Iraqi invasion of Kuwait is given. In Section 3, we introduce the ANN approach. In Section 4, we present an ANN model for Kuwait water consumption time series data from 1980 to 1999, and show that if the delayed variables are identified based on BJ approach, then a better ANN model is obtained than the one based on traditional methods. Some concluding remarks are given in Section 5.

Fig. 1. (a) The time series $x_t$ of the total monthly water consumption in Kuwait in millions of gallons from January 1980 to December 1999. (b) The square root of the time series presented in (a). The solid lines and curves in both graphs represent the fitted trend lines and curves, respectively.
2. Box–Jenkins approach

Let \( x_t \) denote the average water consumption in Kuwait at month \( t \) during the period from January 1980 to December 1999. Fig. 1(a) represents the time series \( x_t \), where the missing portion in the plot corresponds to the unrecorded observations due to the Iraqi invasion of Kuwait (the raw data presented in Fig. 1(a) are available and can be obtained by addressing the first author). The graph shows an upward trend along with seasonal variation whose size is slightly increasing with time. The standard BJ analysis (see [17,18]) involves taking a transformation of the data followed by seasonal and nonseasonal differences to make the data stationary. A stationary time series has a random fluctuation with constant variation around a constant mean. Among all the different kinds of transformations that we tried, a square root function was found to be the best transformation to deal with the seasonality in this data. Fig. 1(b) shows that \( y_t = \sqrt{x_t} \) has seasonal variation with approximately constant size, i.e., an additive seasonality.

Since the main objective of this study is to model the whole time series, \( x_t \), therefore the missing or unrecorded data during the Iraqi invasion of Kuwait must first be predicted. The data from January 1980 to April 1990 that consist of 124 months are used to predict the missing observations. A special type of seasonal autoregressive integrated moving average (SARIMA) model is fitted. The SARIMA model, of order \((1,0,0) \times (1,1,1)_{12}\) in the usual notation of Box et al. [17], fitted to the water consumption data with \( T = 124 \) is given by

\[
(1 - \hat{\phi}_1 B)(1 - \hat{\phi}_{12} B^{12})(1 - B^{12})y_t = \hat{\theta}_0 + (1 - \hat{\theta}_{12} B^{12})a_t, \tag{1}
\]

where \( \hat{\phi}_1 = 0.771, \hat{\phi}_{12} = 0.306, \hat{\theta}_0 = 0.269, \) and \( \hat{\theta}_{12} = 0.854 \) are the estimates of the model parameters obtained using the MINITAB package (release 12). \( B \) and \( B^{12} \) are back shift operators defined as

\[
B^k y_t = y_{t-k}
\]

and \( a_t \) is an independent identically distributed zero mean white noise term. Eq. (1) can be written in the following form

\[
y_t = f(y_{t-1}, y_{t-12}, y_{t-13}, y_{t-24}, y_{t-25}, a_t, a_{t-1}). \tag{3}
\]

After backtransforming all forecasts from the model for the squared root data into the original units, the following results were obtained

(a) \( S = 951680 \), the sum of squared residuals up to time \( T \);
(b) \( \hat{\sigma} = (S/(n - p))^{1/2} = 93.87 \), the estimate of residual standard deviation, where \( n \) is the number of effective observations used in fitting the model, and \( p \) is the number of parameters fitted in the model. Thus, when fitting a model to the water consumption series before invasion with \( T = 124 \), the value of \( n \) is \( 124 - 12 = 112 \), since 12 observations are ‘lost’ by seasonal differencing;
(c) \( \text{AIC} = n \ln(S/n) + 2p = 1021.32 \), the Akaike information criterion;
(d) \( \text{BIC} = n \ln(S/n) + p + p \ln(n) = 1036.19 \), the Bayesian information criterion;
(e) \( \text{SBC} = n \ln(S/n) + p \ln(n) = 1032.19 \), the Schwarz’s Bayesian information criterion;
(f) \( R^2 = 1 - (S/SST) = 0.982334 \), the coefficient of determination where SST is the corrected mean sum of squares of the series.
The statistics AIC, BIC and SBC are model selection criteria. Thus, the minimization of AIC or BIC or SBC is more satisfactory for choosing the “best” model from candidate models having different numbers of parameters. The first term “\( n \ln(S/n) \)” is a measure of “lack of fit” and the remainder is a penalty for increasing the number of model parameters. A more detailed discussion about these criteria and a comparison between them is given by Priestley [19].

Diagnostic checking does not reveal any inadequacies in the model. Thus, the model obtained is used to predict the missing values due to the Iraqi invasion with base = 124 and lead = 24. Fig. 2 depicts the complete time series after predicting the missing values and joining them with the time series presented in Fig. 1(b).

Since one of the main goals of this work is to construct a neural network model for the total monthly water consumption in Kuwait from January 1980 to December 1999, therefore, the BJ approach is used to discover which lagged or delayed variables are necessary to present as inputs to the neural networks. Using the BJ technique on the complete time series presented in Fig. 2, the best model found is SARIMA model of order \((1,1,1)\cdot(1,1,1)_{12}\) given by

\[
(1 - \hat{\phi}_1 B)(1 - \hat{\phi}_{12} B^{12})(1 - B)(1 - B^{12})y_t = (1 - \hat{\theta}_1 B)(1 - \hat{\theta}_{12} B^{12}) \epsilon_t,
\]

where \(\hat{\phi}_1 = 0.659, \hat{\phi}_{12} = 0.354, \hat{\theta}_1 = 0.858,\) and \(\hat{\theta}_{12} = 0.879.\) Alternatively, \(y_t\) can be expressed as follows

\[
y_t = f(y_{t-1}, y_{t-2}, y_{t-12}, y_{t-13}, y_{t-14}, y_{t-24}, y_{t-25}, y_{t-26}, \epsilon_t, \epsilon_{t-1}, \epsilon_{t-12}, \epsilon_{t-14}).
\]

This model necessitates the use of eight input nodes in the input layer for the independent variables in the function \(f\), and one output node in the output layer.

### 3. Neural networks approach

ANN are composed of many nodes that operate in parallel, and communicate with each other through connecting synapses [13–15]. The greatest advantage of a neural network is its ability to model complex nonlinear relationship without a priori assumptions of the nature of the relationship. A multilayer feedforward neural network (MLP) which consists of an input layer, two
hidden layers, each with nonlinear sigmoid function, and an output layer with linear transfer function is used (see Fig. 3). In Fig. 3, the connection between two nodes $i$ and $j$ is characterized by a weight $w_{ij}$. Each node operates by multiplying each incoming signal or input $y_i$ by the weight $w_{ij}$, and then summing up the weighted input. Each hidden layer node performs a single nonlinear transformation of its input

$$z_j = g \left( \sum_i y_i w_{ij} - \beta_j \right),$$

where $\beta_j$ is the bias of the $j$th node, and $g(\cdot)$ is a sigmoidal function given by

$$g(x) = \frac{1}{1 + e^{-x}}.$$

This function belongs to the class of sigmoidal functions and has advantages characteristics such as being continuous, differentiable at all points, and monotonically increasing. It also accepts inputs varying from $-\infty$ to $\infty$ and produces outputs over a finite range from 0 to 1.

Training a network is an essential factor for the success of the neural networks. Among the several learning algorithms available, back-propagation has been the most popular, most widely implemented learning algorithm of all neural networks paradigms. Among the advantages of back-propagation is its ability to store numbers of patterns that exceeds its built-in vector dimensionality. It is based on a multilayered, feedforward topology, with supervised learning. That is, the network is trained in what response it makes to each input it receives. The weights in a

Fig. 3. A neural network architecture that consists of an eight-node input layer, two hidden layers with 12 nodes each and one-node output layer.
network are adjusted by comparing the actual response with the target response in such a way to minimize the sum-squared error, sse, of the network which is given by

\[
\text{sse} = \frac{1}{2} \sum_{p} \sum_{k} (\tilde{z}_k - z_k)^2,
\]

where \(z_k\) and \(\tilde{z}_k\) are the true and predicted output vector of the \(k\)th output node. The \(p\) subscript refers to the specific input vector pattern used. The weights leading into an output node \(k\) are adjusted in proportion to the difference between the actual node output and its target output using Levenberg–Marquardt variation of Newton’s method [16].

4. Neural networks models

4.1. A classical method

A classical method to reconstruct an attractor from a time series by using a set of time delayed samples of the series have been demonstrated by Packard et al. [11]. Using this method, the number of nodes in the input layer is equal to the number of delays or lagged variables \([y_{t-\tau}, y_{t-2\tau}, \ldots, y_{t-k\tau}]\), where \(\tau\) is a time delay, and \(k\) is the number of chosen delays. The output, \(y_{t+p}\), is the predicted value of a time series defined as

\[
y_{t+p} = f(y_{t}, y_{t-\tau}, y_{t-2\tau}, \ldots, y_{t-k\tau}),
\]

where \(P\) is a prediction time into the future.

In this case, we gradually increased the number of lagged variables from 1 to 12 through exhaustive numerical simulation runs. The result is that the predictive capability of the network did not increase if more than four lagged variables were chosen. In fact, increasing the lagged variables past four slows down the convergence rate due to the increase of nodes required in both the input layer and hidden layers. Therefore, the best neural networks architecture found consists of four layers: a five-node input layer, two hidden layers with ten nodes each and one-node output layer. The five input nodes in the first layer are the five delayed values, \(y_t, y_{t-12}, y_{t-24}, y_{t-36},\) and \(y_{t-48}\). The output node \(y_{t+12}\) is the prediction value at \(t + 12\) (i.e., \(\tau = 12, k = 5,\) and \(P = 12\) in Eq. (9)). The network was trained on 168 sets of scaled vectors \((y_{t-48}, y_{t-36}, y_{t-24}, y_{t-12}, y_t, y_{t+12})\) (scaling was done by dividing by the maximum value) consisting of monthly water consumption from 1980 to 1998. Once the network has been successfully trained, it is then used for predicting the 1999 monthly water consumption. Figs. 4 and 5 depict the percent relative error for the training and testing data sets, respectively. The average relative error defined as

\[
\text{ARE} = \frac{100}{n} \sum_{i=1}^{n} \frac{|y_i - z_i|}{y_i},
\]

for the training data and the testing data were approximately 0.137% and 3.92%, respectively. The neural network model can be represented by

\[
y_{t+12} = \sum_{l=1}^{10} w_l^{(3)} g\left(\sum_{j=1}^{10} w_{lj}^{(2)} g\left(\sum_{k=1}^{5} w_{jk}^{(1)} Y_k - \beta_j^{(1)}\right) - \beta_j^{(2)}\right) - \beta_j^{(3)},
\]
where $W^{(1)}$, $W^{(2)}$, and $W^{(3)}$ are the weight matrices for synapses connecting the input nodes with nodes of the first hidden layer, nodes of the first hidden layer to nodes of the second hidden layer, and nodes of the second hidden layer to the output layer, respectively. $Y$ is the input vector that consists of the five delayed values and $g$ is the transfer function given by Eq. (7).

4.2. A new method

A new method to determine the number of nodes in the input layer is described. The method is based on the BJ analysis. Unlike the previous method where the number of delays $m$ is chosen either in an ad hoc basis or from traditional methods, the delayed variables obtained from the BJ analysis are the most important variables to be used as input nodes in the input layer of the neural networks. For the total monthly Kuwait water consumption time series presented in Fig. 2(b), the output $y_t$ can be defined as

$$y_t = f(y_{t-1}, y_{t-2}, y_{t-12}, y_{t-13}, y_{t-14}, y_{t-24}, y_{t-25}, y_{t-26})$$

(12)
The best neural networks architecture found consists of four layers: an eight-node input layer, two hidden layers with 12 nodes each and one-node output layer (see Fig. 3). The nodes in the input layer consist of the lagged or delayed variables, $y_{t-1}$, $y_{t-2}$, $y_{t-12}$, $y_{t-13}$, $y_{t-14}$, $y_{t-24}$, $y_{t-25}$, and $y_{t-26}$, obtained from the BJ analysis presented in Section 2. The output layer node consists of the prediction value at the next month. The number of nodes in the first and second hidden layer was slowly varied from 1 to 12 nodes until a global minimum of $\text{sse} = 1 \times 10^{-4}$ is reached. A set of scaled points was used during the training procedure. Upon convergence, the network is used to predict one month at a time the 1999 monthly water consumption that was not included during the training phase. Figs. 6 and 7 depict the percent relative error for the training and testing data sets, respectively. The average relative error for the training and testing data sets were approximately 0.104% and 2.98%, respectively. The neural networks model can be represented by

$$y_t = \sum_{j=1}^{12} w_j^{(3)} g \left( \sum_{j=1}^{12} w_{jj}^{(2)} g \left( \sum_{k=1}^{8} w_{jk}^{(1)} y_k - \beta_j^{(1)} \right) - \beta_j^{(2)} \right) - \beta_j^{(3)},$$

(13)

Fig. 6. Percentage of relative error between the actual monthly water consumption and the ANN water consumption estimates from March 1982 to December 1998 using training data from 1980 to 1998 in the new method.

Fig. 7. Percentage of relative error between the actual monthly water consumption and the ANN estimates for the testing data from January 1999 to December 1999 using the new method.
where

\[
\begin{bmatrix}
-3.102 & 0.279 & -4.832 & -12.091 & -1.533 & -9.146 & 3.759 & 1.279 \\
0.420 & 0.079 & -1.533 & -5.300 & 0.400 & -6.420 & -1.827 & -0.789 \\
-6.046 & 6.094 & -3.858 & -3.31 & 4.825 & 0.742 & 5.037 & -8.736 \\
-0.798 & -7.837 & -6.247 & 4.265 & -2.817 & 0.578 & 4.515 & 6.186 \\
5.744 & -6.261 & -1.357 & -1.051 & 0.476 & -9.177 & -0.017 & -4.201 \\
\end{bmatrix}
\]

\( w^{(1)} = \) (14)

\[
\begin{bmatrix}
3.233 & -3.710 & 3.095 & 0.241 & -1.408 & -0.443 & 2.200 & -0.963 & -0.530 & -3.258 & 4.851 & 1.704 \\
9.057 & -0.977 & -1.718 & 2.514 & 2.078 & -0.391 & -2.692 & 5.497 & 2.870 & -4.310 & -0.975 & 4.024 \\
-5.376 & -3.490 & 5.751 & 2.720 & -0.357 & 5.318 & -9.304 & -2.644 & -3.577 & 6.281 & 0.955 & 0.421 \\
0.313 & -3.564 & 4.781 & 0.651 & -8.416 & -1.710 & -0.449 & -2.807 & 1.716 & 1.384 & 2.204 & -2.311 \\
-3.439 & -2.264 & -1.612 & 0.809 & 0.137 & -4.021 & 1.278 & -1.996 & 1.450 & 9.556 & -0.696 & -2.562 \\
-1.805 & 3.003 & 1.133 & -2.520 & 0.237 & -2.082 & 1.626 & -3.049 & -0.525 & 2.384 & 1.515 & -2.662 \\
\end{bmatrix}
\]

\( w^{(2)} = \) (15)

\[
\begin{bmatrix}
11.92 \\
1.557 \\
-2.194 \\
-6.926 \\
6.044 \\
-4.446 \\
-3.980 \\
-1.051 \\
1.096 \\
-3.115 \\
2.619 \\
0.949 \\
\end{bmatrix} = \begin{bmatrix}
-9.522 \\
5.359 \\
-1.65 \\
-11.398 \\
-4.289 \\
2.876 \\
-0.928 \\
3.950 \\
3.842 \\
4.679 \\
3.758 \\
-0.23 \\
\end{bmatrix}
\]

\( \beta^{(1)} = \begin{bmatrix}
11.92 \\
1.557 \\
-2.194 \\
-6.926 \\
6.044 \\
-4.446 \\
-3.980 \\
-1.051 \\
1.096 \\
-3.115 \\
2.619 \\
0.949 \\
\end{bmatrix}, \quad \beta^{(2)} = \begin{bmatrix}
-9.522 \\
5.359 \\
-1.65 \\
-11.398 \\
-4.289 \\
2.876 \\
-0.928 \\
3.950 \\
3.842 \\
4.679 \\
3.758 \\
-0.23 \\
\end{bmatrix}, \quad \beta^{(3)} = [1.085], \quad \text{and} \quad Y = \begin{bmatrix}
y_{t-1} \\
y_{t-2} \\
y_{t-12} \\
y_{t-13} \\
y_{t-14} \\
y_{t-24} \\
y_{t-25} \\
y_{t-26} \\
\end{bmatrix}
\]

(17)

is the input vector that consists of the eight delayed variables and \( g \) is the transfer function given by Eq. (7).

Fig. 8 compares the predicted monthly water consumption between the classical method and the new method. The average relative error of the predicted water consumption for the classical
method and the new method after back transforming the data to its original unit were approximately 3.98% and 2.99%, respectively. Based on the average relative errors one can conclude that the combination of BJ approach and ANN (i.e., the new method) is superior in predicting the water consumption than the ANN alone (i.e., the classical method).

5. Concluding remarks

ANN in conjunction with BJ approach have been demonstrated to model the monthly water consumption in Kuwait. The BJ approach was first used to predict the missing values of the monthly water consumption due to the Iraqi invasion of Kuwait. Once the unrecorded monthly water consumption was predicted, BJ approach was then used with the task of discovering the appropriate lagged variables or input nodes in the input layer of the neural networks. This approach presents a superior, and reliable alternative to traditional methods when choosing the appropriate number of delays or lagged variables from a time series. It is found that when the variables of the input layer in ANN is chosen based on the BJ approach rather than on traditional methods, the average relative error for the training and testing data sets are reduced by 24%.

References


Fig. 8. The predicted water consumption for 1999, in millions of gallons, using the classical and the new methods compared with the actual water consumption.