Time Series Forecasting by Using Box-Jenkins Models

Hazim M. Gorgess
Raghad Ibrahim
Dept. of Mathematics/College of Education for Pure Science(Ibn AL-Haitham) /
University of Baghdad

Received in:10 October 2012 , Accepted in:9 December 2012

Abstract

In this paper we introduce a brief review about Box-Jenkins models. The acronym ARIMA stands for “autoregressive integrated moving average”. It is a good method to forecast for stationary and non stationary time series. According to the data which obtained from Baghdad Water Authority, we are modelling two series, the first one about pure water consumption and the second about the number of participants. Then we determine an optimal model by depending on choosing minimum MSE as criterion.

Key Words : Forecasting, Box-Jenkins, autoregressive integrated moving average (ARIMA), Autoregressive (AR), moving average (MA), autocorrelation function (ACF), partial autocorrelation function (PACF)

Introduction

Time series forecasting is one of the most common forecasting techniques used in operation management due to its simplicity and ease of usage. Many time series are non stationary so, we cannot use AR, MA or ARMA processes directly. One possible way to make them stationary is to apply differencing. The first differences namely $(X_t - X_{t-1}) = (1 - B)X_t$, may themselves be differenced to give second differences and so on. The d th differences may be written as $(1 - B)^d X_t$. If original data series is differenced d times before fitting an ARMA(p,q) process, then the model for the original undifferenced series is said to be an ARIMA(p,d,q) process where the letter I refers to integrated and d denotes the number of differences taken.

Autoregressive (AR) Processes

A time series $\{X_t\}$ is said to be an autoregressive process of order p (abbreviated AR(p)) if it is a weighted linear sum of the past p values plus a random shock so that:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + a_t \qquad \ldots(1)$$

where $\{a_t\}$ denotes a purely random process with zero mean and constant variance σ_a^2. Using the backward shift operator "B" such that $B^j X_t = X_{t-j}$, the AR(p) model may be written in the compact form

$$\phi(B)X_t = a_t \qquad \ldots(2)$$

where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$ is a polynomial in B of order p.

An AR process is stationary provided that the roots of $\phi(B) = 0$ lie outside the unit circle, [1].

The simplest case of an AR process is the first order case AR(1) process given by:

$$X_t = \phi_1 X_{t-1} + a_t \qquad \ldots(3)$$

or

$$(1 - \phi_1 B)X_t = a_t \qquad \ldots(4)$$

In order to be stationary the root of $\phi(B) = (1 - \phi_1 B) = 0$ must be outside of the unit circle.

That is for stationary process, we have $|\phi_1| < 1$. The AR(1) process is some times called the Markov process because the value of x_t is completely determined by the knowledge of X_{t-1}.

The autocovariances can be obtained as follows:

$$\gamma_k = E(X_{t-k}X_t) = E(\phi_1 X_{t-1} X_{t-k}) + E(X_{t-k}\, a_t)$$

$$= \phi\gamma_{k-1}, \ k \geq 1 \qquad \ldots(5)$$

and hence the autocorrelation function becomes

$$\rho_k = \phi\rho_{k-1} = \phi_1^k \ , k \geq 1 \qquad \ldots(6)$$

where we use the fact that $\rho_0 = 1$.

The magnitude of these autocorrelations decrease exponentially. For higher order stationary AR processes, the ACF will typically be a mixture of terms which decrease exponentially or of damped sine or cosine waves. For an AR(1) process, the PACF from (6) is:

$$\phi_{kk} = \begin{cases} \rho_1 = \phi_1 & \text{for } k = 1 \\ 0 & \text{for } k \geq 2 \end{cases} \qquad \ldots(7)$$

Hence, the PACF of AR(1) process cuts off at lag 1. A useful property of an AR(p) process is that it can be shown that the PACF is zero at all lags greater than p. This means that the sample PACF can be used to help determine the order of an AR process (assuming the order is unknown as it is usually the case) by looking for the lag value at which the sample

PACF cuts off (meaning that it should be approximately zero, or at least not significantly different from zero, for higher lags).

Moving Average (MA) Processes

A time series $\{X_t\}$ is said to be **a moving average process** of order q (abbreviated MA(q)) if it is a weighted linear sum of the last q random shocks so that:

$$X_t = a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q} \qquad \ldots(8)$$

or

$$X_t = \theta(B)\, a_t \qquad \ldots(9)$$

where $\theta(B) = (1 - \theta_1 B - \ldots - \theta_q B^q)$ is a polynomial in B of order q.
It can be shown that a finite order MA process is stationary for all parameter values. However, it is customary to impose a condition on the parameter values of an MA model called the "invertibility condition". This moving average process is invertible if the roots of $\theta(B) = 0$ lie outside of the unit circle, [2]. Moving average processes are useful in describing phenomena in which events produce an immediate effect that only lasts for short periods of time. To discuss further properties of the MA(q) process, let us consider the following simpler case when $\theta(B) = 1 - \theta_1 B$ we have the first order moving average MA(1) process

$$X_t = a_t - \theta_1 a_{t-1} = (1 - \theta_1 B)\, a_t \qquad \ldots(10)$$

where $\{a_t\}$ is a zero mean white noise process with constant variance σ_a^2. It can be shown that the autocovariances of the process are:

$$\gamma_k = E_{X_t\, X_{t-k}} = \begin{cases} (1+\theta_1^2)\sigma_a^2 & k=0 \\ -\theta_1\sigma_a^2 & k=1 \\ 0 & k>1 \end{cases} \qquad \ldots(11)$$

and hence the autocorrelation function becomes:

$$\rho_k = \begin{cases} 1 & k=0 \\ \dfrac{-\theta_1}{1+\theta_1^2} & k=1 \\ 0 & k>1 \end{cases} \qquad \ldots(12)$$

which cuts off after lag 1.
Using (11) and (12) the PACF of an MA(1) process can be easily seen to be:

$$\phi_{11} = \rho_1 = \frac{-\theta_1}{1+\theta_1^2} = \frac{-\theta_1(1-\theta_1^2)}{1-\theta_1^4}$$

$$\phi_{22} = -\frac{\rho_1^2}{1-\rho_1^2} = \frac{-\theta_1^2}{1+\theta_1^2+\theta_1^4} = \frac{-\theta_1^2(1-\theta_1^2)}{1-\theta_1^6}$$

In general:

$$\phi_{kk} = \frac{-\theta_1^k(1-\theta_1^2)}{1-\theta_1^{2(k+1)}}, \text{ for } k \geq 1 \qquad \ldots(13)$$

The PACF of an MA(1) model tails off exponentially in one of two forms depending on the sign of θ_1 (hence on the sign of ρ_1). If alternating in sign, it begins with positive value, otherwise, it decays on the negative sign.
For the general qth order moving average process the variance is:

$$\gamma_0 = \sigma^2 a \sum_{j=0}^{q} \theta_j^2 \qquad \ldots(14)$$

where $\theta_0 = 1$, and the other covariances are

$$\gamma_k = \begin{cases} \sigma_a^2(-\theta_k + \theta_1\theta_{k-1} + \ldots + \theta_{q-k}\theta_q) & k = 1,2,\ldots,q \\ 0 & k > q \end{cases} \qquad \ldots(15)$$

Hence the ACF becomes

$$\rho_k = \begin{cases} \dfrac{-\theta_k + \theta_1\theta_{k-1} + \ldots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \ldots + \theta_q^2} & k = 1,2,\ldots,q \\ 0 & k > q \end{cases} \qquad \ldots(16)$$

Thus the ACF cuts off at lag q. This property may be used to try to assess the order of the process (i.e. the value of q) by looking for the lag beyond which the sample ACF is not significantly different from zero. The PACF of the general MA(q) process tails off as a mixture of exponential decays and / or damped sine waves depending on the nature of the roots of θ(B) = 0. The PACF will contain damped sine waves if some of the roots are complete. The dual relationship between an AR(p) and an MA(q) processes exists and may be described as follows [3]. A finite order stationary AR(p) process corresponds to an infinite order MA process and a finite order invertible MA(q) process corresponds to an infinite order AR process, this relationship also exists in the ACF and PACF. The AR(p) process has its autocorrelations tailing off and partial autocorrelations cutting off while MA(q) process has its autocorrelations cutting off and partial autocorrelations tailing off.

Table (1) represent the quantity of water consumption of Baghdad City for the period (2000-2008) and table (2) represent the number of participants for the period (2000-2008), where figure (1) represent series water consumption and figure (2) represents the series of number of participants which are non stationary. It is clear that the first is stationary.

Modeling by Box-Jenkins Method

Box-Jenkins is a famous method which has a high qualified in time series analysis that reflect behaviour of series while if it was seasonal or non seasonal. In this research ,we modeled two series. We depend on SPSS and Minitab programmes to determine tables and figures. Table (3) represents the ARIMA model for water series consumption and table (4) represents ARIMA models for series number of participants.

Conclusion

It is clear that by applying Box-Jenkins methods depending on minimum MSE as criterion to choose an optimal model, the best method for water series consumption is (2,0,1) which has no differences and means the series stationary, and the best model for series number of participants is (2,1,2).

References

1. Chris Chatfield (2000) Time Series Forecasting, Univ. of Bath, Dep. Of Math. Sciences.
2. Wei,W.W.S. (1989) Time Series Analysis, Addision Wesley Publishing Company Inc.
3. Margherita G. (2010) ARIMA and SARIMA Models.

Table (1):Quantity of Water Consumption of Baghdad City for the period (2000-2008)

period	year	Water Consumption
31/12/1999 – 30/4/2000	2000	150199 058
1/5/2000 – 31/8/2000		117836291
1/9/2000 – 31/12/2000		110120423
31/12/2000 – 30/4/2001	2001	192430197
1/5/2001 – 31/8/2001		260990152
1/9/2001 – 30/12/2001		248181835
31/12/2001 – 30/4/2002	2002	192892689
1/5/2002 – 31/8/2002		198365209
1/9/2002 – 30/12/2002		195628949
31/12/2002 – 30/4/2003	2003	112386361
1/5/2003 – 31/8/2003		110389561
1/9/2003 – 30/12/2003		111387961
31/12/2003 – 30/4/2004	2004	106390698
1/5/2004 – 31/8/2004		168258446
1/9/2004 – 30/12/2004		119950103
31/12/2004 – 30/4/2005	2005	100917650
1/5/2005 – 31/8/2005		134876155
1/9/2005 – 30/12/2005		140109023
31/12/2005 – 30/4/2006	2006	112130164
1/5/2006 – 31/8/2006		168329582
1/9/2006 – 30/12/2006		113161258
31/12/2006 – 30/4/2007	2007	164912005
1/5/2007 – 31/8/2007		112033870
1/9/2009 – 30/12/2007		148864908
31/12/2007 – 30/4/2008	2008	202032553
1/5/2008 – 31/8/2008		223789575
1/9/2008 – 30/12/2008		129457565

Table (2):Number of Participants for the period (2000-2008)

period	year	Number of Participants
31/12/1999 – 30/4/2000	2000	552825
1/5/2000 – 31/8/2000		564314
1/9/2000 – 30/12/2000		556327
31/12/2000 – 30/4/2001	2001	571845
1/5/2001 – 31/8/2001		571731
1/9/2001 – 30/12/2001		555949
31/12/2001 – 30/4/2002	2002	557170
1/5/2002 – 31/8/2002		557109
1/9/2002 – 30/12/2002		557140
31/12/2002 – 30/4/2003	2003	557102
1/5/2003 – 31/8/2003		555611
1/9/2003 – 30/12/2003		556357
31/12/2003 – 30/4/2004	2004	558600
1/5/2004 – 31/8/2004		559519
1/9/2004 – 30/12/2004		559837
31/12/2004 – 30/4/2005	2005	560903
1/5/2005 – 31/8/2005		561551
1/9/2005 – 30/12/2005		563386
31/12/2005 – 30/4/2006	2006	566547
1/5/2006 – 31/8/2006		568128
1/9/2006 – 30/12/2006		571265
31/12/2006 – 30/4/2007	2007	573615
1/5/2007 – 31/8/2007		575894
1/9/2009 – 30/12/2007		576591
31/12/2007 – 30/4/2008	2008	578429
1/5/2008 – 31/8/2008		580834
1/9/2008 – 30/12/2008		582955

Table (3):ARIMA Model for Water Series Consumption

ARIMA	Parameters	MSE
(1,0,0)	p(1) 0.2218	0.07942
(0,1,0)	no autoreg. or movi. average	–
(0,0,1)	q(1) – 0.1980	0.07977
(1,1,0)	p(1) – 0.4183	0.11435
(1,0,1)	p(1) 0.2044 q(1) – 0.0185	0.08303
(0,1,1)	q(1) 0.9142	0.09364
(1,1,1)	p(1) 0.2026 q(1) 0.9161	0.09405
(2,0,0)	p(1) 0.2245 p(2) – 0.0108	0.08302
(0,2,0)	no autoreg. or movi. average	–
(0,0,2)	q(1) – 0.2502 q(2) – 0.0993	0.08250
(2,2,0)	p(1) – 0.8795 p(2) – 0.3389	0.21035
(2,0,2)	p(1) – 0.8198 p(2) – 0.4803 q(1) – 1.1620 q(2) – 0.8325	0.08240
(0,2,2)	q(1) 1.4100 q(2) – 0.4599	0.12312
(2,2,2)	p(1) – 0.9689 p(2) 0.0336 q(1) 0.0079 q(2) 0.8890	0.12964
(1,0,2)	p(1) 0.7532 q(1) 0.6492 q(2) 0.2550	0.08358
(1,2,0)	p(1) – 0.6493	0.22316
(2,0,1)	p(1) – 0.8216 p(2) 0.1759 q(1) – 0.9322	0.07903*
(2,1,0)	p(1) – 0.5041 p(2) – 0.2087	0.11492
(2,1,1)	p(1) – 1.1045 p(2) – 0.1028 q(1) – 0.9768	0.10921
(1,2,1)	p(1) – 0.1770 q(1) 1.1350	0.11346
(1,1,2)	p(1) 0.2120 q(1) 0.9152 q(2) 0.0229	0.09728
(2,1,2)	p(1) 0.7569 p(2) – 0.2899 q(1) 1.2970 q(2) – 0.2021	0.09458

Table (4):ARIMA Models for Series Number of Participants

ARIMA	Parameters	MSE
(0,1,0)	no autoregr. or movin. average	–
(0,0,1)	q(1) – 0.9241	0.00010437
(1,1,0)	p(1) – 0.3621	0.00008927
(1,0,1)	cannot be estimated with these data	–
(0,1,1)	q(1) 0.4067	0.00008747
(1,1,1)	p(1) – 0.0440 q(1) 0.3684	0.00009159
(2,0,0)	p(1) 0.7202 p(2) 0.2791	0.00009805
(0,2,0)	no autoregr. or movin. average	–
(0,0,2)	q(1) – 0.5492 q(2) – 0.5263	0.00013817
(2,2,0)	p(1) – 0.9207 p(2) – 0.7206	0.00009393
(2,0,2)	p(1) 1.3055 p(2) – 0.3062 q(1) 0.6447 q(2) – 0.2022	0.00010666
(0,2,2)	cannot be estimated with these data	–
(2,2,2)	p(1) – 0.4740 p(2) – 0.5481 q(1) 0.4746 q(2) – 0.3658	0.00012311
(1,0,2)	cannot be estimated with these data	–
(1,2,0)	p(1) – 0.6501	0.00016255
(2,0,1)	p(1) 0.8022 p(2) 0.1971 q(1) 0.1456	0.00010206
(2,1,0)	p(1) – 0.4047 p(2) – 0.2017	0.00009108
(2,1,1)	p(1) 0.5321 p(2) 0.2363 q(1) 1.1833	0.00006983
(1,2,1)	cannot be estimated with these data	–
(1,1,2)	p(1) – 0.5343 q(1) – 0.4204 q(2) 0.5306	0.00007867
(2,1,2)	p(1) – 0.9599 p(2) – 0.4592 q(1) – 0.8780 q(2) 0.0874	0.00006792*

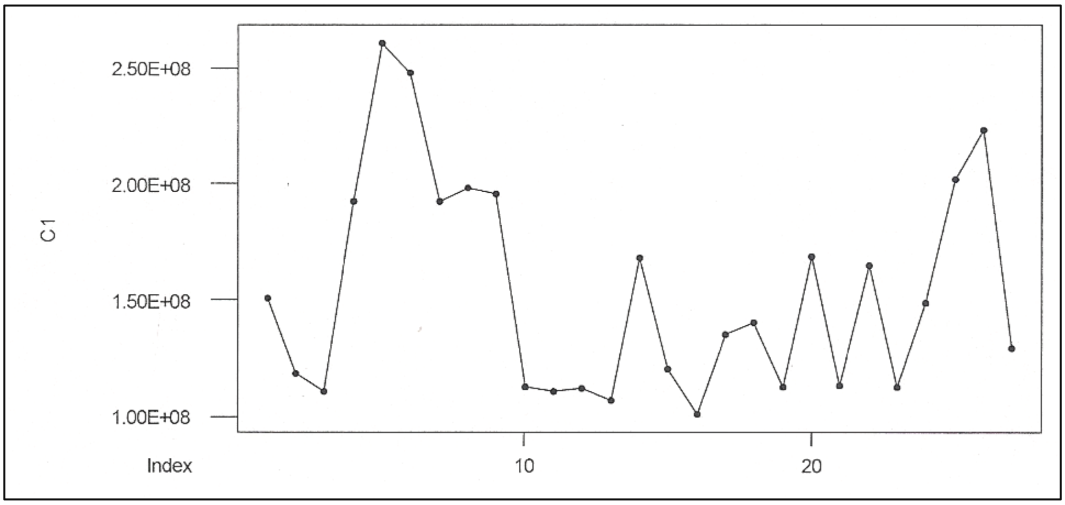

Fig (1):Series Water Consumption

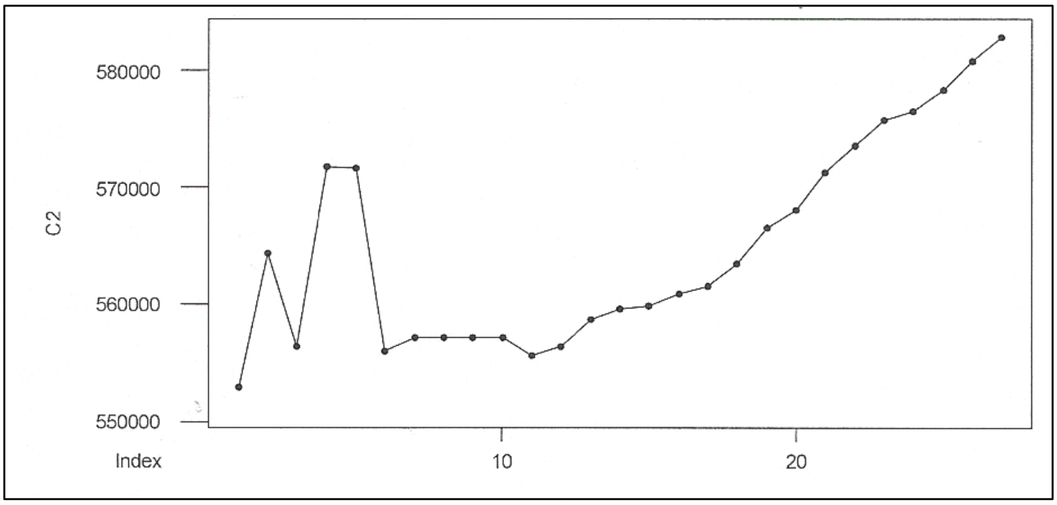

Fig (2):Series Number of Participants

التنبؤ بالسلاسل الزمنية بإستخدام اساليب بوكس جينكيز

حازم منصور كوركيس
رغد إبراهيم
قسم علوم الرياضيات / كلية التربية للعلوم الصرفة (ابن الهيثم) / جامعة بغداد

استلم البحث في :10 تشرين الاول 2012 ، قبل البحث في :9 كانون الاول 2012

الخلاصة

في هذا البحث قدمنا نبذة مختصرة عن نماذج البوكس-جينكيز تمثل ARIMA اختصاراً بالانحدار الذاتي والاوساط المتحركة وهو اسلوب جيد للتنبؤ بالسلاسل الزمنية المستقرة وغير المستقرة. ووفقاً للبيانات التي حصلنا عليها من دائرة ماء بغداد قمنا بنمذجة سلسلتين: الاولى سلسلة استهلاك الماء والاخرى سلسلة اعداد المشتركين ومن ثم تحديد الانموذج الافضل بالاعتماد على اختيار أقل متوسط مربعات الخطأ معياراً لاختيار الانموذج.

الكلمات المفتاحية : التنبؤ، بوكس جينكيز، النماذج المختلطة، انحدار ذاتي، اوساط متحركة، دالة الارتباط المشترك، دالة الارتباط المشترك الجزئي الذاتي.