International Journal of Statistical Sciences *ISSN 1683–5603*
Vol. 5, 2006, pp 59-72

Short-Term Forecasting of Influent Chemical Oxygen Demand

Faisal Hossain
Department of Civil and Environmental Engineering, Box 5015
Tennessee Technological University
Cookeville, TN 38505-0001
Email: fhossain@tntech.edu

Wun Jern Ng and Say Leong Ong
Wastewater Biotreatment Group, Department of Civil Engineering
The National University of Singapore
10 Kent Ridge Crescent, Singapore 119260

[Received January 6, 2004; Revised June 29, 2005; Accepted July 7, 2006]

Abstract

A simplified methodology is developed to determine the order of a Box-Jenkins (ARMA - Auto Regressive Moving Average) time-series model for forecasting the influent Chemical Oxygen Demand (COD) for a wastewater treatment plant (a Sequencing Batch Reactor-SBR). The determination of the order was based on temporal correlation analyses. The Auto Regressive (AR) and the Moving Average (MA) processes were investigated separately and then combined over the modeling period to identify the processes that resulted in better forecasts. A purely AutoRegressive model of order 2 was found to predict better within reasonable limits the influent COD with a 10 day lead. The finding agreed well with previous research reports that state that the moving average component is insignificant for describing the temporal dynamics of influent characteristics for treatment plants. The Box-Jenkins methodology used herein, being marginally theoretical in formulation, has appeal to plant managers as a quick fix and short-term (10 day lead) forecasting method of influent wastewater characteristics.

Keywords and Phrases: Statistical Analysis; Auto Regressive and Moving Average; Box-Jenkin's Methodology; Sequence Box Reactor (SBR).

AMS Classification: 37M$_{xx}$

1 Introduction

The ability to foresee the consequences of actions and events can form a very important aspect of control in engineering. From the dawn of recorded history, and probably before, man has sought to forecast the future. In wastewater treatment plants, forecasting of influent characteristics can be useful as it helps in determining the future operational criteria. For example, a forecasted surge in the strength of influent wastewater as measured by the COD (Chemical Oxygen Demand) can help the plant manager revise the mean cell residence times (i.e., the average time that waste is treated for) by implementing reduced wasting of sludge. A higher COD is an indicator of more treatment time required by the microbial organisms contained in the sludge of the plant. Similarly, if the pH (pouvoir Hydrogen) of the influent is forecast to be unusually high or low then this can possibly send signals for pH control to prevent damage to the biomass of the treatment system.

A question however remains: *How accurate can the forecasts be made on a short-term basis by a quick fix method (i.e., simple enough to be implemented rapidly by the plant manager) so that they may be acceptable to be considered for operations planning?* This paper addresses this question. It considers a simplified method for identification of a Box-Jenkins (ARMA- AutoRegresive Moving Average) model for forecasting the influent COD of a wastewater treatment plant - an SBR (Sequencing Batch Reactor) - treating high strength industrial effluent. The method is comprised of identification of the following in a step by step manner:

1) the Auto Regressive and Moving Average processes for influent COD variation.

2) the best ARMA model that yields the highest accuracy in forecasting during the modeling range.

3) the validation of the ARMA model and its forecasting accuracy with a 10 day lead.

4) ways of improving forecasting accuracy of the simplified Box-Jenkins approach.

Figure 1 shows the temporal variation of the INFCOD (Influent COD) of the SBR during the assessment period which spanned a total of 60 days (total period including validation comprised 105 days). The dates have been referenced from the date of the start of the study. The range DAY 1 - DAY 60 was used for evaluation and model identification (steps 1 and 2). The subsequent range of DAY 61 - DAY 105 (not shown in Figure 1) was used for model validation (step 3). The statistical package - SAS (Statistical Analysis System), version 6.12 for Window 95^{TM} - was used for the statistical computation on a Intel Pentium 166 MHz PC. Since the Box-Jenkins methodology requires stationarity, the time series of influent COD was detrended (i.e., differenced) in a preprocessing analysis (as evident in Figure 1).

2 Box and Jenkins (ARIMA) Methodology

The Box-Jenkins (1976) ARMA (Auto Regressive Moving Average) methodology has been used for time-series analysis because it is a powerful yet simple methodology for forecasting a uni-variate time-series (Brocklebank and Dickey, 1986). Generally, Box and Jenkins methodology has the following merits:

1) The concept is easy to understand and the model is simple to formulate (Box and Jenkins, 1976).

2) It has become very popular and is the most common methodology used by modellers to for time series analysis of influent characteristics (Akaike and Nagawa, 1972; Nakamura and Akaike, 1981; Hiraoka and Fujiwara, 1992; Tao et al., 1994; Sales et al., 1994).

2.1 Autoregressive Process

The general pth order autoregressive process for a variable x (to be forecast) is defined by

$$x_t - \phi_1^+ x_{t-1} - \phi_2^+ x_{t-2} - \cdots\cdots\cdots - \phi_p^+ x_{t-p} = \varepsilon_t \tag{1}$$

where $\phi_1, \phi_2 \cdots\cdots \phi_p$ are constants and ε_t is the random error (of the forecast) more appropriately known as *white noise*. The model is denoted by $AR(p)$. The white noise element is normally distributed with zero mean and variance σ^2. This type of process of process (*Equation. 1*) is known as the autoregressive process, since it represents a regression of x_t on x_{t-1}.

Equation (1) can be rewritten as,

$$x_t - \sum_{i=1}^{p} \phi_i x_{t-i} = \phi_0 + \varepsilon_t \tag{2}$$

or,

$$(1 - A(L))x_t = \phi_0 + \varepsilon_t \tag{3}$$

or,

$$x_t = \frac{\phi_0}{(1 - A(L))} + \frac{\varepsilon_t}{(1 - A(L))} \tag{4}$$

where $A(L) = \sum_{i=1}^{p} \phi_i L^i$ is the lag polynomial and L is the lag operator defined as $L^k x_t = x_{t-k}$.

For a simple $AR(1)$ model, (*Equation 1*) can be rewritten as,

$$x_t = \mu + \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \cdots\cdots + \phi^{t-1}\varepsilon_1 + \phi^t(x_0 - \mu) \tag{5}$$

where the mean (expected) value of x_t is μ.

The knowledge of ϕ helps in forecasting future values of x. By statistical manipulation it can be shown that, a forecast l steps into the future is,

$$\mu + \phi^1(x_t - \mu) \tag{6}$$

with error,

$$\varepsilon_{t+l} + \phi\varepsilon_{t+l-1} + \cdots\cdots\cdots + \phi^{l-1}\varepsilon_{t+1} \tag{7}$$

Solution to the $AR(1) : x_t = \phi_0 + \phi_1 x_{t-1} + \varepsilon_t$ model becomes (after repeated subsitution with an initial condition x_0)

$$x_t = \phi_0 \sum_{i=0}^{t-1} \phi_1^i + \phi_1^t x_0 + \sum_{i=0}^{t-1} \phi_1^i \varepsilon_{t-i} \tag{8}$$

Taking the expected value of equation (8), we obtain

$$Ex_t = \phi_0 \sum_{i=0}^{t-1} \phi_1^i + \phi_1^t x_0 \tag{9}$$

Updating by l period, me obtain

$$Ex_{t+s} = \phi_0 \sum_{j=0}^{t+l-1} \phi_1^j + \phi_1^{t+l} x_0 \tag{10}$$

For large sufficiency large values of t both equations (9) and (10) will converge to a constant (μ) since $|\phi_1| < 1$ (the stationarity condition for the AR(1) process). However, there is no reason to believe that the dependence of x_t on past values should be limited to the previous observation x_{t-l} only.

A simple way to determine the order of the Autoregressive process is described in the next section (Figure 2).

2.2 Moving Average Process

As an example of another stochastic process based on the difference of values between the present and the past (i.e. correlated errors), the moving average (MA) model of order q is presented as MA(q),

$$x_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots\cdots\cdots \theta_q\varepsilon_{t-q} \tag{11}$$

where $\theta_1, \theta_2, \cdots\cdots \theta_q$ are constants and ε_t is white noise.

Using the mean μ, x_t can be rephrased for an MA(1) process as,

$$x_t = \mu + \varepsilon_t - \theta\varepsilon_{t-1} \tag{12}$$

Just as in the case for the autoregressive process, a future value of x can be predicted l steps ahead into the future as,

$$x_{t+l} = \mu - \theta^l(\varepsilon_{t-1}) \tag{13}$$

with error,

$$\varepsilon_{n+l} + \theta\varepsilon_{n+l-1} + \cdots\cdots\cdots + \theta^{l-1}\varepsilon_{n+1} \tag{14}$$

The determination of the order of the moving average component has been described in the next section (Figure 2).

3 ARMA Model (Box and Jenkins)

The most natural generalization of the ARMA model is the combination of the AR and the MA process as described in the previous section (Gilchrist, 1976). Thus, if a first-order MA process is combined with a first-order AR process, the resultant model is known as a *mixed autoregressive and moving average process*(ARMA) of the order $(1, 1)$ as shown below.

$$x_t - \phi x_{t-1} = \varepsilon_t - \theta\varepsilon_{t-1} \tag{15}$$

Using the mean μ, the general ARMA model of order (p, q) is given by,

$$(x_t - \mu) - \phi_1(x_{t-1} - \mu) - \cdots\cdots\phi_p(x_{t-p} - \mu) = \varepsilon_t - \theta_1\varepsilon_{t-1} - \cdots\cdots - \theta_q\varepsilon_{t-q} \tag{16}$$

3.1 A Simplified Methodology for ARMA Model Identification

The primary task of the time series analysis using the Box and Jenkins (ARMA) approach is to identify the order of the ARMA process. In order to identify the order of the model the following approach was adopted as indicated in the flowchart below (Figure 2). To maintain simplicity in the approach, the R^2 measure was used to identify the order that was most representative of the time-series. However, such measures are not always reliable for time series modeling and other criterion, such as the Akaike Information Criterion (AIC- Akaike, 1972), are more robust measures under certain cases.

3.1.1 The Auto Regressive Process for INFCOD

The X as in Figure 2 refers to the influent COD (INFCOD) variable to be analyzed in the time series. Linear regressions (with intercept) were performed of INFCOD on lagged values of INFCOD upto lag 9. This means that the following models were tried,

$$
\begin{aligned}
INFCOD_t &= A + B * INFCOD_{t-1} \\
INFCOD_t &= A + B * INFCOD_{t-1} + C * INFCOD_{t-2} \\
INFCOD_t &= A + B * INFCOD_{t-1} + C * INFCOD_{t-2} + D * INFCOD_{t-3} \\
INFCOD_t &= A + B * INFCOD_{t-1} + \cdots\cdots\cdots\cdots\cdots + J * INFCOD_{t-9}
\end{aligned}
$$

where $A, B, C \cdots$ are constants.

Table 1 shows the values for $R2(R^2)$ and *Adjusted R2 (Adj.*R^2*)* for the models. We have also considered the use of $Adj.R^2$ (see Notations) to minimize the effect of over-fitting.

Table 1 reveals that an AR model of order $p = 3$ is the best fit for the autoregressive process as it provides the optimum $R2$ value against $Adj.R^2$ value.

As a next step, the data available for the Modelling (fitting) period (Day 1 to Day 60) was divided into four subsequent parts, each of an arbitrary 20 days. 20 days were devised based on the amount of data available. Each of these parts were used to fit (and calibrate) the AR model and then forecast for the subsequent 10 days, as,

Modelling Period	Forecasting Period
DAY 1 - 20	DAY 21 - 30
DAY 11 - 30	DAY 31 - 40
DAY 21 - 40	DAY 41 - 50
DAY 31- 50	DAY 51 - 60

AR models upto order $p = 3$ were fitted, i.e., AR(1), AR(2) and AR(3). For each AR model, the hypothesis of *white noise* was tested using the *chi-squared* statistic for residuals (Box and Jenkins, 1976; Brocklebank and Dickey, 1986). For all the three models the residuals were found to be statistically significant in constituting a *white noise* series. The parameter coefficients ϕ_1, ϕ_2 and ϕ_3 were all less than unity and the AR models were of the form as in *Equation 5*. The mean used for forecasting was the arithmetic mean of INFCOD. Judging from the $R2$ value and the average deviation (error) in absolute terms, the AR(2) model was found to forecast best among the three. Hence, p for the ARMA model was decided as $p = 2$. Figure 3 shows how well the AR(2) model fitted in the modeling period based on the above concept of a moving base period for calibration.

3.1.2 Moving Average Process

In a similar manner to AR, linear regressions (with intercept) were performed of INFCOD on lagged errors $\varepsilon_{t-1}, \varepsilon_{t-2}, \cdots\cdots \varepsilon_{t-9}$. Table 2 shows the $R2$ values for the above nine models. As seen from the table, both the $R2$ and Adjusted $R2$ show consistent increase with increasing lagged errors. Since, the MA model would be too unwieldy if too many lagged errors are involved, the Adjusted $R2$ as percentage of $R2$ was taken as the criteria for estimating the optimum value of q for the MA process. q was therefore estimated to be 3.

In the next step, the data available for the modelling (fitting) period (D) was found statistically inadequate to describe the INFCOD time series based on the MA process. Hence it was conjectured that the MA process was insignificant (i.e., $q = 0$) in describing the temporal dynamics of INFCOD. This observation agrees well with that of Nakamura and Akaike (1981) who failed to achieve better forecast by incorporating the MA process in their influent time series model for a power plant.

Next, the ARMA models of order (2,1), (2,2) and (2,3) were fitted in the Modelling period (DAY 1 - 60) in a similar fashion as for the AR or MA process. The ARMA models were each tested for the *white noise* hypothesis using the *Chi-squared statistic*, and were found to be statistically significant. Since both the $R2$ and mean residual values of the ARMA models were found to be inferior to the AR(2) model, the Moving Average component of the ARMA process was discarded. Table 3 show the $R2$, average deviation (error) in absolute terms and the % error of the model fit for the ARMA(2,1), ARMA(2,2), ARMA(2,3) and the AR models. The final model selected was therefore an AR(2) model which had been identified as the most significant. This model was used to forecast INFCOD values in the validation range of DAY 61 - 105.

3.2 The Identified ARMA Model

The final time series model was therefore an AutoRegressive model of order 2 (AR(2)) of the form

$$INFCOD_t = \mu + \phi_1(INFCOD_{t-1} - \mu) + \phi_2(INFCOD_{t-2} - \mu) + \varepsilon_t \qquad (17)$$

where ε_t represents the *white* noise series.
The model in *Equation 17* was fitted in the Modelling range of DAY 1 -60, as,

$$INFCOD_t = 2497 + 0.51(INFCOD_{t-1} - 2497) + 0.36(INFCOD_{t-2} - 2497) + \varepsilon_t (18)$$

where INFCOD is the COD of the influent measured in mg/l.

The *Chi-squared* values for ε_t and tests of significance are shown in Table 4. As it can be seen, the residuals are statistically significant at 0.05 level as a random error series. Figure 4 shows the frequency distribution of the residuals ε, indicating an almost normal distribution.

3.3 ARMA Model Validation

With the final time series model AR (2) properly formulated, the model (Equation.18) was then validated using data ranging from DAY 61 to DAY 105. A moving window period of 60 days was used for calibration of the model and the forecast was done for the subsequent 10 days. Figure 5 shows the validation of the AR(2) model. It is found

that the AR(2) model yields an average absolute error of 873 mg/l (Figure.5). The moving window period for calibration of the AR(2) model is shown below.

Calibration (Moving window period)	**Forecasting**
DAY 1 - DAY 60	DAY 61 - DAY 70
DAY 11 - DAY 70	DAY 71 - DAY 80
DAY 21 - DAY 80	DAY 81 - DAY 90
DAY 31 - DAY 90	DAY 91 - DAY 100
DAY 71 - DAY 130	DAY 101 - DAY 105

4 Discussion

Sales et al. (1994) and Tao et al. (1994) have performed similar forecasting analyses using the Box-Jenkins AR model for predicting inflows for a catchment. However, there appears to be little literature for comparison of the forecast of influent strength of wastewater (COD) using the Box-Jenkins model for a SBR wastewater treatment plant. Existing literature has shown that records of around 1000 data points are necessary for identifying a versatile time series model and forecasting analysis for

wastewater treatment control (Hiraoka and Fujiwara, 1992). In this example of forecasting analysis, only 120 data points have been used in evaluating the series due to lack of further data, a very common and realistic scenario in small-scale wastewater treatment plants. The average error in predicting the influent COD with a 10 day lead time (the usual lead time required by the plant manager to assess options for treatment control) had been about 873 mg/l (approximately 35% error). Therefore, it is hypothesized that if more data from the SBR plant can be available, the time-series and forecasting analysis should allow more accurate with forecast errors being within a more tolerable range ($>$ 20% error). With more data spanning a greater length of time, more temporal features of COD variation like "seasonality" and "trend" could be more accurately identified and thus enhance the quality of the Box-Jenkins analysis. If a reasonably well forecasting sequence is possible, then the plant operator may use the forecasting analysis to estimate his operational criteria 10 days ahead. Furthermore, the simple algorithmic nature of model identification shown in Figure 2 can be translated conveniently to a computer routine for computation of the ARMA model automatically from on on-line data monitoring system at the plant site. As most treatment plants of today are being monitored on-line for the influent, effluent and process. characteristics (Andrews, 1992), the computerization of this approach is therefore an attractive option. However, it is recognized that there may be other important statistical criteria that are not examined here. For example, more robust and sophisticated methods for evaluating forecast models exist - such as split sample predictive power tests of the initial models considered (.e.g AR vs ARMA) and the use of Ljung-Box Q statistics for a more definitive test of white noise. Nevertheless, we hope that this paper will lead to further studies involving a wider range of methods with the objective of designing a simplified procedure for the forecasting of influent COD. Such procedures will have appeal for plant managers and practitioners alike.

Statistical Notations

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

Where SST = Total Sum of squares (model sum of squares)

SSR = Sum of squares due to regression (also called explained sum of squares)

SSE = Sum of squares due to error (also called unexplained sum of squares).

$$Adjusted\ Rsquare\ (Adj.R^2) = 1 - (1 - Rsquare)\ \frac{N - K}{N - K - 1}$$

N = No. of data points , K = No of regression terms in the model

References

[1] Akaike, H. and Nagagawa, T. (1992). 'Statistical Analysis and Control of Dynamic Systems'. Tokyo, Japan: Kluwer Academic Publisher.

[2] Andrews, J. F. (1992). 'Dynamics and Control of the Activated Sludge Process'. Water Quality Management Library, Vol. 6, Lancaster, USA: Technomic Publishing Co. Inc.

[3] Brocklebank, J. C. and Dickey, D. A. (1986). 'SAS System for Forecasting Time Series'. (1986 edition), Cary, USA: SAS Institute.

[4] Box, G. E. P. and Jenkins, G. M. (1976). 'Time Series Analysis: Forecasting and Control'. Revised Edition, Oakland, CA, USA: Holden -Day.

[5] Gilchrist, W. (1976). 'Statistical Forecasting'. Bath, UK: John Wiley and Sons.

[6] Hiraoka, M. and Fujiwara, T. (1992). 'The Use of Time-series Analysis in Hierarchical Control Systems'. In Dynamics and Control of the Activated Sludge Process, Vol. 6, Water Quality Management Library, ed by J. F. Andrews, Lancaster: Lewis Publishers, 133 - 167.

[7] Nakamura, H. and Akaike, H. (1981). 'Statistical Identification for Optimal Control of Supercritical Thermal Power Plant'. Automatica, 17, 1, 143-155.

[8] Sales, P. R., Pereira H. D. B. and Vieira, A. M. (1994). 'Linear Procedures for Time-series Analysis in Hydrology'. In Stochastic and Statistical Methods in Hydrology and Environmental Engineering, The Netherlands: Kluwer Academic Publishers, 3, 105 - 117.

[9] Tao, T., Corbu, I., Penn, R., Benzaquen, F. and Lai, L. (1994). 'Seeking User input in Inflow Forecasting'. In Time Series Analysis in Hydrology and Environmental Engineering, Netherlands: Kluwer Academic Publishers, 3, 99 - 104.